**NHL Game Simulation**

**BACKROUND AND METHODS**

For our simulation, we looked at data from the National Hockey League (NHL) for the current season (2017-18). This included multiple variables for each game played thus far, but for our purposes we cleaned the original data to focus on only the following variables: Date, Home Team, Away Team, Home Goals, Away Goals, and the end result of the game. We then created probabilities (described below).Traditionally, NHL Standings (by points) are scored as followed: two points for a win, one point for an overtime win, and one point for a Shootout loss. For simplicity, we chose to disregard Overtime and Shootout points and simply scored each win as a point. Our primary goal was to simulate the outcome of the remaining games left in the 2017-18 season, and to predict the which team is likely to be the top ranked team overall (by points) at end of the regular season. There are multiple ways to predict the result of a hockey game, but for this project we chose to use expected goal values as the determining factor of each game outcome. “Expected goals is a mere tool that assists in better describing the likelihood that past events occurred and predicting the likely course of a side’s future performance…All expected goals models incorporate major parameters that most strongly influence the likelihood that a chance will turn into a goal.” (<https://www.pinnacle.com/en/betting-articles/Soccer/expected-goals-and-big-chances/QFU2JE3AQPK8VQV7>). These factors include: the distance of the attempt from goal, and type of attempt made (i.e. a regular shot or header). According to the Pinnacle article referenced above, the expected goal value is also useful because it recognizes and accounts for the fact that “not all shots are created equally…which led to the idea that a side with a higher expected goals tally than their opponents is more likely to outscore them.” Here, expected goals are calculated according to the method posted by Mark Taylor in “the Power of Goals”. (<http://thepowerofgoals.blogspot.co.uk/2013/10/finishing-and-hitting-target-in-mls.html>) Taylor’s method uses “zones” which are related to the location of where any given attempted shot was taken. It looks at these zones of each shot taken in a game along with shot accuracy, goals, and blocked shots.

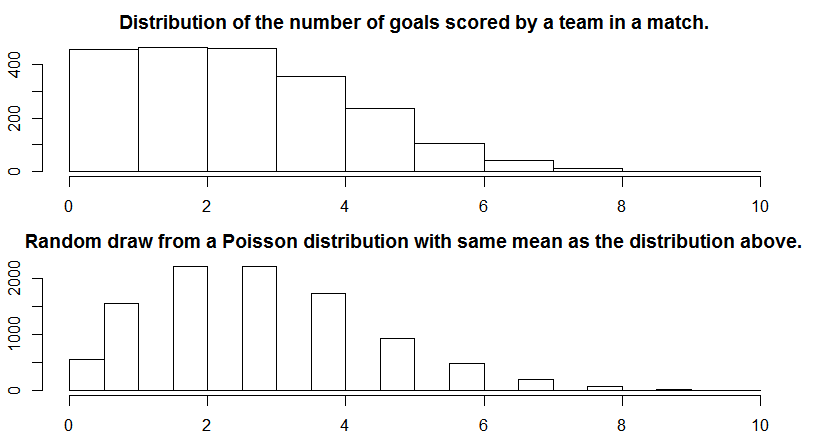
To get the expected goals scored values for our simulation, we first brought in the table for all the games that haven’t been played yet. Then, we created a column called “HG” which is the average home goals scored overall and a column called “AG” which is the average away goals scored overall. Both HG and AG are average goals scored across the entire league, which accounts for the home and away effects. “TG” is the average goals scored value, regardless of whether the goals were made at home or away. Next we calculated the average goals scored and average goals conceded specific to both the home and away team in a given game using a right join. Last, we used all the previous columns to calculate the expected goals using the below calculation:

**× × ×**

**SIMULATION AND MODELING**

We used two methods in our project: a Monte Carlo simulation and a Poisson Distribution. The Poisson Distribution is a discrete probability distribution that “expresses the probability of a given number of events occurring in a fixed interval of time.” The Poisson Distribution can be applied when: “(1) the event is something that can be counted in whole numbers; (2) occurrences are independent, (3) the average frequency of occurrence for the time period in question is known; and (4) it is possible to count how many events have occurred”. (<https://www.umass.edu/wsp/resources/poisson/>) (<https://en.wikipedia.org/wiki/Poisson_distribution>) The Poisson Distribution is commonly used to predict outcomes of sport games so we chose to use this as one of our methods in our project. As noted above, it can be used to count events (in our case events are goals scored) within a specific time period (one hockey game). While there are three periods in a hockey game, for our purposes we considered each single game to be an independent event. We used the number of average expected goal values to determine any given game’s outcome. To get an idea of the Poisson Distribution would be a good fit for our simulation, we modeled the match results for average goals scored and compared that to a model produced by using a Poission Distribution to estimate the likely average amount of goals scored per game (see below).

***Modeling Average Goals Scored per Game: Actual Goals Scored vs. Poisson Distribution Model Goals***



As shown above, a quick comparison between the actual distribution of the average number of goals scored in a game and the modeled Poisson distribution of likely number of average goals scored in a game shows relatively similar results. When comparing the two charts, the actual average number of goals scored relatively closely follows the modeled Poisson distribution. In both graphs, it is less likely to score a higher amount of goals per game than the average goal total, which is around two or three goals per game.

Next, we created probabilities for each of the 31 teams in the league. To do this, we used a Monte Carlo simulation was used to create “what if” scenarios. The Monte Carlo simulation uses the “rpoisson” function to randomly pick a score for the home and away team in a given game by using the the expected goal value for both the home and away team as a rate. In this case, the rate used was the lamba. In a Poisson Distribution, the Lambda is “a parameter that captures the average number of events in an interval.” (<https://math.stackexchange.com/>) In our case, it made sense to use the expected goals scored value as our Lambda, since it represented the average number of an event (goals) in an interval (a single hockey game). Each game was run 500 times and the results were put into a new data frame (temp.df). Finally, we calculated the probabilities by averaging these results.

**RESULTS**

While this simulation has only modeled regular season games, we can now see which team has the highest probability of be ranked first going into the playoffs. From our simulation, it looks like the Tampa Bay Lightning are the current favorite at a rate of 83.4%. At this point in the actual season, Tampa Bay has the best overall record of 48 wins and 18 losses. Using our model, we can also look at all each individual team's probabilities to a respective ranking out of 31 total teams.

**IMPROVEMENTS TO MODEL**

There are a few improvements we felt could be made to our model going forward. It would have been ideal to score the team rankings (determined by points) more accurately by including Overtime and Shootout data. It appears that the probabilities are heavily influenced by the current overall win and loss records at this point in the season.

Additionally, this simulation was built with the assumption that each hockey game is an independent event. In reality, a sports game is not really an “independent event” and can be affected by many outside factors. For example, any given game could be influenced by the team’s record of the last few games (i.e. a 5 game losing streak), team rivalries, traveling distance to the game, personal issues, etc. It would be hard to prove any game is not at least in some way either positively or negatively affected by previous games.