# Final Analysis

# of

# Data Topic

By

Kilian Hammersmith

Source: <https://moneypuck.com/moneypuck/playerData/seasonSummary/2021/regular/skaters.csv>

Submission Date (05/15/2022)

Fundamentals of Data Science and Analytics - CS345

Professors: Dr. Adam R. Albina, Dr. Stephen Shea

## Introduction

The goal of this analysis is to determine how shot quality relates to goal scoring success based off National Hockey League data, individual skater advanced statistics from MoneyPuck for the 2021-2022 season. Players are becoming progressively more skilled and game strategy is starting to reflect the change in player ability. Highly skilled players tend to hold on to the puck longer to create a better scoring opportunity, rather than settle for a lower quality shot. The classic idea in hockey is that all shots are good and shot volume is what is important. For example, corsi percentage, which is a popular advanced statistics in hockey, considers the differential between shots for the opponent and shots by the players team. There is no consideration for shot quality. I intend to develop a model that displays the importance of shot quality and the relationship it has to a players expected goal total. The model I create will be a multiple regression with the variables that best predict a players expected goal totals.

The exploratory data analysis I ran on the data was revealing about what variables impact a players expected goal values. Shot quality absolutely matters when considering what a player’s production can be expected to be. High quality shots showed a clear correlation, as expected. Medium quality shots showed a much higher correlation that what might be expected, even higher than high quality shots. Low quality shots showed little correlation. The more low-quality shots that a player gets does not increase the expected goals. Both corsi and fenwick percentage both had little correlation to expected goals for an individual player. The percentage of shots a player gets blocked also has some correlation and will be considered for the model. Shot quality variables will certainly be a part of the model. I will use the value per shift in order to have each player equally represented, given a large distribution of playing time. A final goal of the analysis is to display the relationship between quality and quantity of shots, followed by a multiple regression model that will predict a players expected goals based on shot quality.

## Model Development

The analysis done during the exploratory phase led to a clear opportunity to create a multiple regression model. The model will include advanced statistics relating to the shot quality and quantity of an individual. This is the most effective way to evaluate the weight of each variable and create a predictive value based on a player’s shot inputs. Metrics like corsi and fenwick are current advanced ways that shots are evaluated; however, they are rather limited because they account for total volume in shooting. Shot volume does not always lead to increased success, as seen in the exploratory data analysis when low quality shots were plotted against expected goals. To create the model, I started with many variables: high quality shots per shift, medium quality shots per shift, percentage of shots that are not blocked, the percentage of shot attempts that end up on goal, low quality shots per shift, total shots per shift, corsi, and fenwick, see line 142. I used a backward selection technique referencing how the adjusted R-squared value react to removing different variables, see line 139-184. The model that had the best results in terms of R-squared and adjusted R-squared included high-quality shots per shift, medium quality shots per shift, low quality shots per shift, and total shots per shift. The model with coefficients is below:

xGoalsPerShift = .309HighQPerShift + .133MedQPerShift + .024LowQPerShift + .002TShotsPerShift -.0004

I then moved to testing the assumptions made in a linear model. First, I used the plot function to plot the model, see line 198. I then looked through the plots to verify that they behave as they should for a model like this. The first plot evaluated was the Residuals vs Fitted.

There is no pattern in errors given the line is about zero across all fitted values. The next consideration is the Quantile-Quantile graph to evaluate if errors are normally distributed.

For most values, the errors seem to be mostly normally distributed. The model appears to be sufficiently linear. A histogram of the residuals is also relatively normal distributed, see line 199.

I also used a cross validation technique to test the model. I first split the data into a training and test set. A model was created using the training set. Then a prediction was created using the test data set. To compare the data, a mean of the difference between the prediction data and the actual data set was calculated. Then a correlation was ran comparing the two values. The correlation coefficient was .986, which is very significant. This sufficiently validates the model, see lines 184-202.

## Analysis Conclusions

The analysis leading to the creation of this model revealed interesting relationships between shot quality and quantity with respect to goal scoring output, or what should be expected on average. The model results show that each quality of shot that is recorded and total shots are the greatest predictors of expected goal totals. The coefficients indicate that high quality shots are the most significant variable in the model, followed close by medium shot quality. Low quality shots and total shot volume are relevant variables but are not as significant as the other two variables in the model. The model indicates that the more, higher quality and medium quality shots a player, the higher expected goal value they will have. While shot volume and low-quality shots also have a role, they are not nearly as important as the shots of greater quality. A player should try to create medium and higher quality shots rather than settling for lower quality shots but should not avoid taking low quality shots if that is the only option available. The old game play strategy of shot volume is king, in line with a corsi evaluation, is clearly not as effective as looking to create better scoring chances as often as possible. A variable I did not expected to be so significant was the medium quality shots per shift. This variable had a very strong correlation to expected goals per shift, even higher that high quality shots per shift. His may be an indicator that medium quality shots taken often might be the goal of a player capable of creating those opportunities.

Below are the plots for the model that show linearity. Referenced above, the residuals all lie around zero, the quantile-quantile seems to be rather normally distributed, and the residuals are also normally distributed.

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart

Description automatically generated

Another unexpected result of the model was that blocked shot percentage or percentage of shot attempts that were on net did not fit in with the model. I expected that if a player were to have less shots blocked or hit the net on a higher percentage of their shots, that would be a significant indicator of expected goal scoring success; however, that did not turn out to be the case.

A limitation of the model is that all values for expected goals, and thus the estimation of shot quality, are all estimates. Given the nature of hockey, there is no way to have an exact value for the expected outcome of a potential game play scenario. The model is at the liberty of the data collection and the method of determining the values.

REFERENCES

Jiang, Jinhang. “Linear Regression Analysis in R.” *Medium*, Towards Data Science, 30 May 2021, https://towardsdatascience.com/linear-regression-analysis-in-r-fdd59295d4a8.

Include any libraries with proper attribution to the authors (many authors have preferred citation information listed with the library documentation - search the CRAN repository to find them).

@Article{,

title = {Welcome to the {tidyverse}},

author = {Hadley Wickham and Mara Averick and Jennifer Bryan and

Winston Chang and Lucy D'Agostino McGowan and Romain François and

Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester

and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan

Milton Bache and Kirill Müller and Jeroen Ooms and David Robinson

and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and

Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani},

year = {2019},

journal = {Journal of Open Source Software},

volume = {4},

number = {43},

pages = {1686},

doi = {10.21105/joss.01686},

Brie, Evelyne. “Introduction to Cross-Validation in R.” *RPubs*, https://rpubs.com/evelynebrie/crossvalidation.