CMPT353 D100

Computational Data Science

Project: Predicting NHL Goalie Save Percentage

Professor: Greg Baker

Submitted: August 4, 2023

Aidan Howker - 301446390
Navraj Gosal - 301360678

The Problem

In the NHL, many goalies seem to be prone to dramatic swings in effectiveness from season to season - it feels as though many goalies have great seasons followed by terrible seasons, and vice versa. The most common measure of a goalie's success is his save percentage, or the percentage of shots put on goal by the opposing team that are stopped by the goalie. There are many statistics for goalies that are recorded and posted online (such as goals allowed, number of quality scoring chances faced, etc.), and many that are more difficult to quantify (such as emotions, exact positioning, etc.). Fans around the league often seem to point to the easily-recorded statistics such as save percentage and goals allowed as indicators of a goalie's future success, and yet the mystery of these season-to-season inconsistencies remain.

Our question is as follows: ***can we predict the save percentage of a goalie in the next season, based on his recorded statistics in the previous season? If so, are there any statistics in particular that are strongly correlated to the next season's save percentage?***

We developed a machine learning model based on as many available statistics as possible in order to attempt to answer this question, followed by some other statistical analysis including linear regressions and a Pearson correlation test.

The Data

To find data for the National Hockey League, we looked to a popular hockey data website, MoneyPuck.com. MoneyPuck provides free-to-use statistics from all players and all teams from the 2008-2009 season onwards. The data is organised in simple .csv files, separated by season. We downloaded goaltender statistic files from each of the previous ten seasons and concatenated them into one large document before beginning the cleaning process. The statistics we used that were provided in the csv are as follows:

- Games played
- Icetime (minutes)
- Goals against
- Expected goals against
- Unblocked shot attempts
- Rebounds allowed
- Expected rebounds allowed
- Puck freezes (covering up the puck and drawing a whistle from the referee)
- Expected puck freezes
- Shots on goal (by the opposing teams)
- Expected shots on goal
- Play stoppages
- Expected play stoppages

- Play continuations
- Expected play continuations
- Flurry adjusted expected goals against (adjusted for breakaways/rush chances by the opposing teams)
- Low danger shots
- Medium danger shots
- High danger shots
- Low danger goals
- Medium danger goals
- High danger goals
- Low danger expected goals
- Medium danger expected goals
- High danger expected goals
- Blocked shot attempts (blocked by the goalie's teammates

note: 'expected' stats are what an average goalie would be expected to put up, given the quantity/quality of shots and situations faced by the goalie in question.

Before beginning the cleaning process, we decided to remove goalie-seasons with less than 20 games played. A typical backup goalie in the NHL will play around 20 games per season (with most starting goalies playing 40-60). Goalies who have only played a handful of games in a given season could negatively impact our model, since the small sample sizes provided in their data may be prone to significant variance and could be more easily swayed by potential outlier games. This left us with data for approximately 50-60 different goalies per season.

Despite the data being nicely organised by MoneyPuck.com, there were still many steps required to clean the data and format it properly for our analysis. We began by filtering out the columns that were not necessary for our question. There was a 'situation' column, which added rows for specific situations such as power plays. Since we were more interested in goalies' success overall, we filtered on this column so that only the rows labelled 'all' remained - these are the cumulative stats for the goalie in that given season. After filtering, we removed the column entirely along with other columns containing information we did not need, such as position (all positions were Goalie), team (we were not interested in certain team's goalies), and penalties taken. We then manually created the save percentage column, which was not already included, and used a helper function to create a column for the next year's save percentage - this would be our response variable. Once this was completed, we made sure to remove rows where 'next_sv_pct' was equal to None (these are seasons where the goalie did not play at least 20 games the following season, and therefore didn't have a 'next save percentage'). We then output our DataFrame as another .csv file to be used in our analysis. For clarification, our final data frame contained one row per goalie-season, with the various statistics as columns.

Techniques Used

During our data cleaning, we used Pandas data manipulation techniques learned in this class, including combining data frames, filtering data, and creating/applying custom functions (without loops) to help with our analysis. We also read from and wrote to .csv files.

Once we had our data cleaned, we created a random forest regression machine learning model using scikit-learn, as we did in our weekly exercises. We split the data in order to train the model on one subset of the data, and test it on another subset. After creating our model, we also used linear regressions (also from scikit-learn) and a Pearson test for correlation, the contents of which will be explained in the next section. Plots were created with matplotlib.pyplot.

Results/Findings

We were surprised to discover that our model was unable to effectively predict a goalie's next save percentage in any meaningful way. We tweaked the model's parameters, but even our most effective version of the model was only capable of achieving a 0.399 training score and a validation score of only 0.04. It was essentially unable to predict the next save percentage for a goalie, despite the 30 statistics used in the model (accumulated over a decade of NHL play).

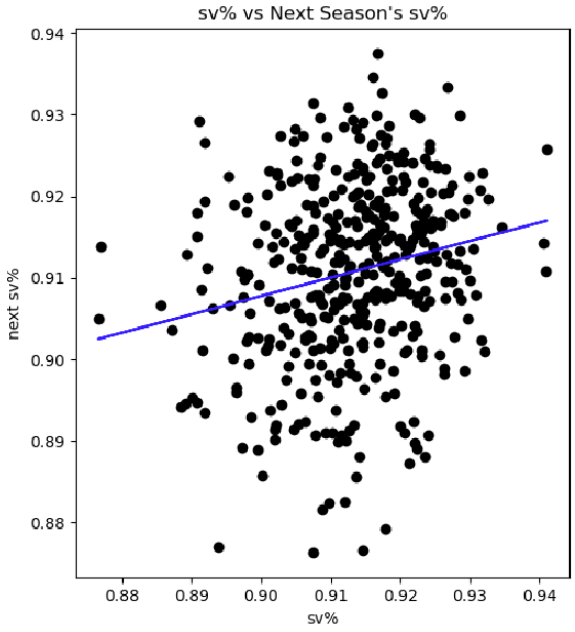
Intrigued, we looked into the variables the model was weighting the most heavily by looking at the feature importances of our completed model. Unsurprisingly, the model found that a goalie's save percentage in a given season was the best predictor for that goalie's save percentage in the following season, with a Gini importance of 29.4%. The number of rebounds allowed and the number of shots blocked by teammates were the second and third most important variables according to the model with Gini importances of 7.1% and 6.4%, respectively. All other variables had Gini importances below 5.5%. One interesting note is that neither goals (1.6%) nor expected goals (2.4%), two variables that many would think have a strong impact on a goalie's future success, contributed to the model in any significant way.

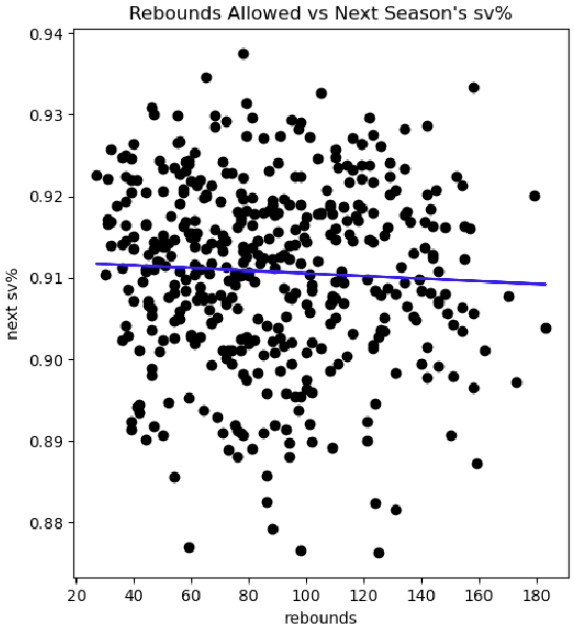
To further confirm our results, we conducted linear regressions on these three most important variables, comparing them to our response variable (the next season's save percentage). Visualisations are provided below, and the results provided the same results as our model - none of these variables were strongly correlated with the next season's save percentage. To further investigate, we also ran a Pearson test for correlation between save percentage and the next year's save percentage. A p-value of 1.62e-05 was far below our alpha value of 0.05, which interestingly suggested that we should reject our null hypothesis of no correlation between the two statistics, and conclude that there indeed was some correlation. The correlation coefficient of 0.21, however, suggested that any correlation that exists is not particularly strong.

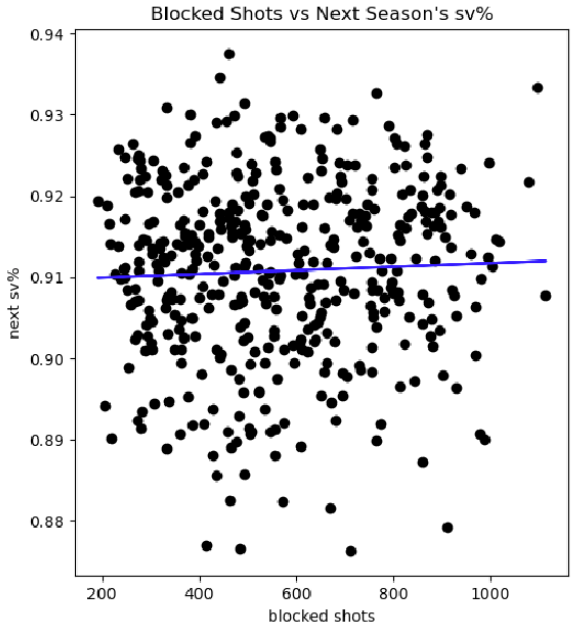
Overall, our results were quite shocking - we had anticipated prior that we would be able to predict a goalie's save percentage to some extent using the 30 statistics included in our model and 10 seasons' worth of data, but we found that no publicly available statistics (nor a model factoring in many of them at once) were able to effectively predict a goalie's save percentage.

Visualisations

Scatterplots for the linear regressions of the three most important features of the model vs the next season's save percentage:







Correlation coefficients for plots:

Save percentage:	0.0426
Rebounds:	0.0025
Blocked shots:	0.0019

Linear regressions found no clear correlation between any of these statistics and the next season's next save percentage. Even a goalie's save percentage, which is the strongest feature from our model, is not a good predictor of their save percentage for the next season.

Problems/Reflection

Although the project ran smoothly overall, there were still a few areas in which improvements could be made. It is possible that with more data, we would have been able to develop a stronger model - perhaps adding more variables from other sources or including more seasons of data could have improved our ability to predict goalies' save percentage. However, most of the statistics that media use on a daily basis were included, so we still believe that we found an interesting result in that none of those statistics proved to be an effective predictor for save percentage.

On top of the other recorded variables we didn't include, there are countless unrecorded or publicly unavailable variables such as a goalie's diet, mentality, confidence, and so on that could factor into their future success.

We also wanted to explore the common phenomenon known as 'hot/cold streaks' for goaltenders: if their previous handful of games proved to be a particularly strong estimation for the next game's success, but when predicting data accumulated over entire seasons proved impossible with our current tools, we decided that goal was a little ambitious, at least for this project.

Accomplishment Statements

Aidan Howker

- Researched different online sources of NHL data to find the source that best suited our project's needs, minimising the data cleaning and formatting required before we began our analysis.
- Cleaned and formatted data downloaded from the internet in Python, using Pandas to filter, modify, create and drop rows/columns of data. This allowed the analysis portion of the project to run smoothly and help us come to a conclusion.
- Helped revise a random forest regression machine learning model using scikit-learn that attempts to predict NHL goalie save percentage based on other publicly available statistics in order to answer our project's primary question (can we predict NHL goalies' save percentage using publicly available statistics?).
- Effectively worked with a partner on a data science project, using Git to collaboratively work on code while also maintaining continuous verbal and written communication with my partner. This resulted in efficient work and also served to reduce the likelihood of any conflicts within the team during the project.
- Wrote a thorough report that presented our problem, the data we used, the analysis we conducted, and our results in such a way that a non-expert in NHL statistics could understand our work.

Navraj Gosal

- Gathered, collated and cleaned information from online sources which resulted in ten years of useful NHL goalie statistics for our defined problem.
- Created a Random Forest machine learning model in Python using the sci-kit learn library in order to predict NHL goalie save percentages based on various statistics.
- Tuned and tested machine learning model by modifying multiple parameters such as tree depth, leaf size to optimize for reliability and accuracy.
- Analyzed variables using the Pearson test to explore correlation between the strongest predictor and our response variable. This uncovered an interesting relationship which provided further insight in building our conclusions.
- Collaborated with a partner on a data science assignment with consistent communication to efficiently complete tasks and allow for swift progress before deadlines.