Shift: A New Era for Professional Hockey

Final Report Document
Jacob Mosseri (1003588405) & Christopher Mazzuca (1003150181)
Word Count: 1996
Penalty: 0%
Revised: 2018-12-02

Introduction

This document will outline and describe a completed project; the attempt to apply machine learning techniques to Professional Ice Hockey (namely, the National Hockey League or NHL). The project, "*Shift*", has been constructed with the aim of using tools learned in class to help ease the decision-making of the administration in the NHL.

Explicitly, the goal of this project is to build a model that predicts the statistics of a currently active player in the NHL for any required season (split into the categories of a "goaltender", and "skater"). With that information, a team can evaluate how effective its players will be in the upcoming year, and if there are more viable options in free agency. The thesis for this project is that when given information such as team, age and games played for a given player, predictions can be made with an acceptable[1] degree of accuracy.

Applying machine learning to this type of prediction-based analysis on players is cutting edge in the sport. The motivation for this project is deeply rooted in the belief that there needs to be a revolution in analytics for not only the NHL, but all professional sports leagues. With a project of this nature, there will of course be many complications. Some of the troubles faced include how the data is processed, the accuracy of the models, and time constraints on the project. All of these obstacles, along with the successes, will be discussed in further detail below.

Software Structure

This section will dive into the specifics around the structure of the software. The software can be broken down into four main components: data collection, data

[1] In sports management, a prediction will not be useful if it bears considerable error and inaccuracy. This will be considered when discussing the results of the project later in the document.

preprocessing, building the model, and optimization. Although these will be presented as separate processes, they are interconnected and took place relatively in conjunction with each other. In Figure 1 below, a detailed flow diagram of the software is shown.

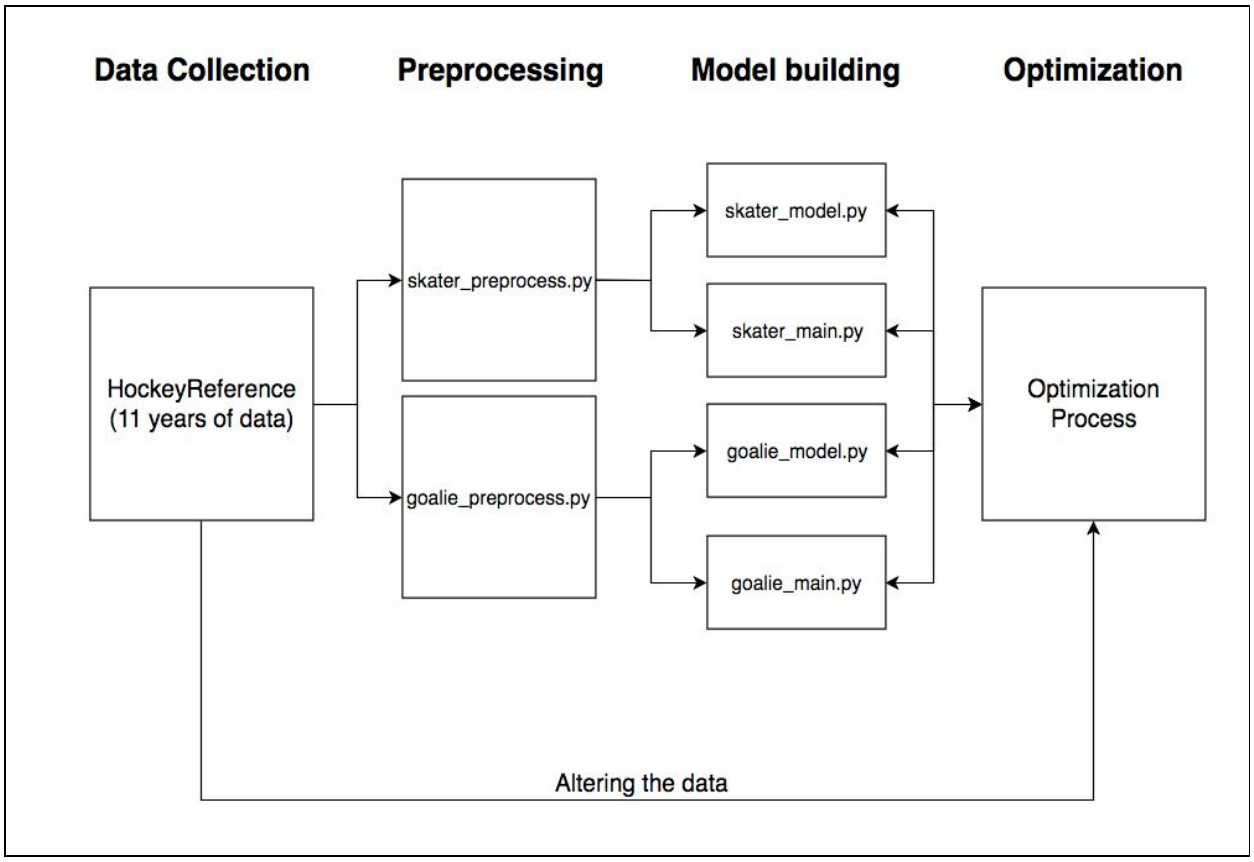

Figure 1: Visual representation of the software structure

The workflow generally moved from data collection and processing into model building. However, there was a continual feedback and feedforward development in the optimization process, hence the multidirectional arrows in the flow chart. This structure is slightly different and is more advanced from the one proposed earlier in the project, due to the many changes and realizations that took place throughout the process.

It was very difficult to develop a concrete plan from the beginning of this project, with little intuition on how the model may respond to different data configurations and processing. As a result, a majority of the development in this project was the product of reading different approaches in various resources

(mainly not related to hockey), trying them, and evaluating the results. The software was designed for such a process, and this is evident in Figure 1 above.

Sources of Data

Data collection and preprocessing was the most important stage of the project, and the one that took up the most time. All of the used data was collected by scraping the Web. The data was mainly comprised from Hockey-Reference.com, an online database that houses years of NHL data, grouped into skaters and goaltenders.

In the Project Proposal, it was noted that several different websites were being considered, including ones with alternative sources of data (i.e. advanced statistics). There were several reasons for using HockeyReference.com exclusively, mainly due to the constraints of time, and because it included unique identifiers for each player (to eliminate the error of two players having the same name, which happens fairly frequently).

From this source, 11 years of NHL statistics were collected and stored in CSV files (for players and goaltenders). These 11 seasons included a "lockout" season, where only 48 games were played instead of the regular 82 (due to Collective Bargaining Agreement in 2012). Along with this obstacle, there was missing data that had to be dealt with. For instance, the statistic of faceoff percentage was missing, if a player never took a faceoff in that season. This data, for simplicity, was removed before preprocessing.

At this point the data was ready to be preprocessed for later use. All of the data was separated into goaltenders and skaters, preprocessed in the files *skater_preprocess.py* and *goalie_preprocess.py*. In this step, the categorical inputs (position for skaters and identifiers for both players and goalies) were one hot encoded and this method of encoding was saved as a pickled dictionary. In addition, both the inputs and the labels were normalized. Due to the labels being normalized, the mean and the standard deviation of our labels was also stored as a NumPy array. Since the mean and the standard deviation remain somewhat constant year over year in the era of hockey that our data is from, it was a viable solution to use the mean and standard deviation from previous years to reverse the normalization process after our models have predicted. The processed data was then split into training and validation and these four NumPy arrays were saved, as

this would be used later on. The features and the labels of our two neural nets can be found below:

Features	Labels
Identifier	Wins
Age	Losses
Games Played	T/O
Games Started	SA
Team	SV
	SO
	GPS
	QS
	RBS
	GSAA

Features	Labels
Identifier	Goals
Age	Assists
Games Played	Points
Position	PS
Team	EVG
TOI	EVA
	PPG
	PPA
	SHG
	SHA
	GWG
	Shots
	Hits
	Blocks

Figures 2 & 3: T-Chart listing the features and labels for goalies (left) and skaters (right)

Models

One model was made for skaters and one model was made for goalies. Both models were composed of 3 linear layers and one dropout layer. This dropout layer was placed in between the first and second linear layers and had a probability of 0.5. The second and the third layers of this linear layer were both activated by a relu function. The reason that we used this linear multilayer perceptron is because MLPs are most suited towards regressionary prediction [5].

Other models that we tried were RNNs for both skaters and goalies and a CNN for skaters. In order to input the data into the RNNs we adjusted the preprocessing to have each player's history over time as the input and then the output would be the next year. For the CNN, we inputted a team's data as a whole as a method of forcing the neural net to recognize the impact of a player's team that season on their final point production.

Interestingly, both of these models had less accuracy than the multilayer perceptron. This could be due to the fact that the layers were able to extrapolate more complicated patterns that the other models weren't able to. As a result, we

decided to move forward with a fully connected linear model, and altering the number of layers.

Training and Optimization

After the models were built, a training loop was built in order to train our models to our data. This training loop consisted of an Adamax optimizer with a learning rate of 0.001 and a weight decay of 0.0001. In the optimization process, we tried many different optimizer and loss functions. We found that the Adamax optimizer produced best, along with a Smooth L1 loss. The Smooth L1 Loss function was used in order to calculate loss due to this ensuring that gradients are steady no matter the size of the difference. Other optimizers that were considered were Adam and Adagrad due to all of them having an adaptive gradient.

The hyperparameters were chosen while trying to optimize the validation accuracy of our model. The accuracy was defined as a confidence interval, where the prediction was defined as correct if it was within 5% of the label. This confidence was chosen because we felt that anything greater than this wouldn't be useful for a practical application in the industry. However, in the optimization process we experimented with different confidence intervals, and noticed that the results were relatively linear. If the interval was increased by 5%, the accuracy increased by 5% respectively.

As you can see in Figures 4 and 5 below, the final validation accuracy for the goaltender model reached 89%, and the final validation accuracy for the skater model reached roughly 49%. Note that these results are better than what was presented in the final presentation, and we will continue to improve these results (out of genuine interest in the project). The main increase in our results came from data preprocessing, and not from changing the optimizer and loss functions. With more time, we would like to try different techniques used in other sports, and see how this affected the results.

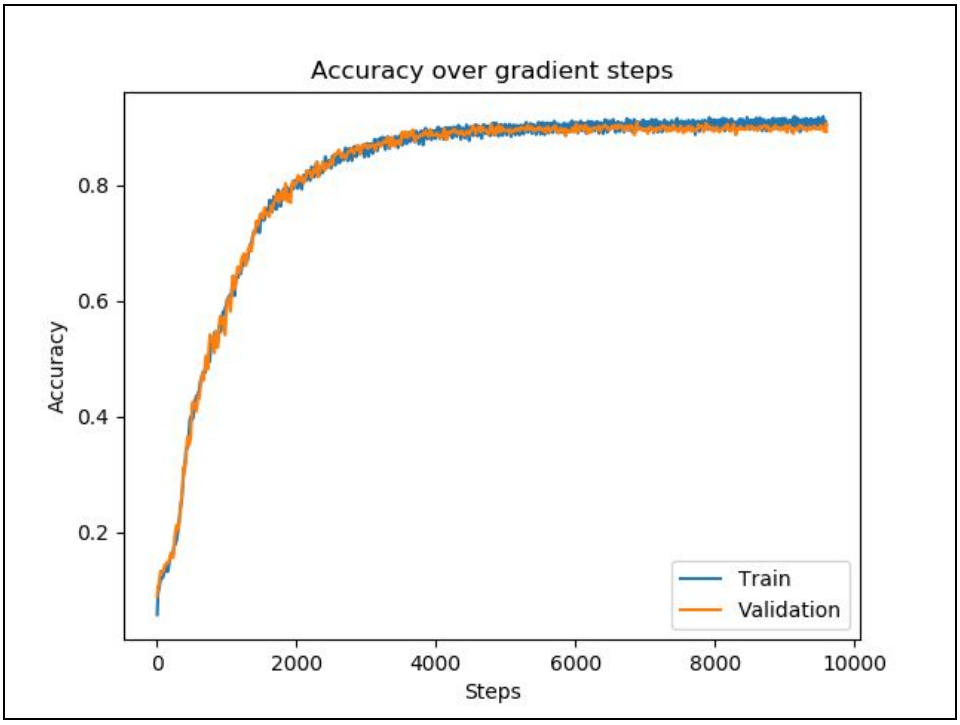

Figure 4: Training and Validation accuracy for the goaltender model

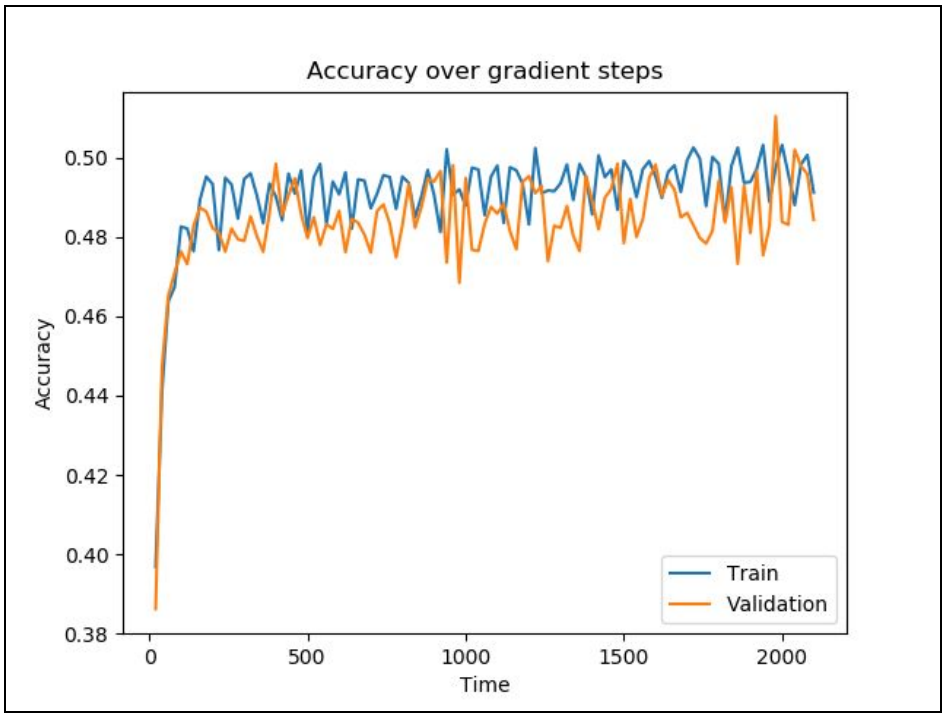

Figure 5: Training and Validation accuracy for the skater model

Potential Ethical Concerns

One notable ethical concern is that this program could change the way players are valued. This would shift the focus on heeding to the certain categories that are subjectively determined hold intrinsic value. Thus, players with different skill sets must be taken into account when approaching this project.

Unconventional hockey players, such as the Washington Capitals' Tom Wilson, the Pittsburgh Penguins' Tom Sestito, and the Boston Bruins' Tommy Wingels, may be undervalued due to the their style of play (less points, but more aggressive in order to protect teammates). Also, hockey players with non-quantifiable hockey skills such as Chicago's Jonathan Toews, Los Angeles' Jonathan Quick, and Las Vegas' Jonathan Marchessault (exhibiting qualities such as leadership) may also become more undervalued.

It is important to consider that the statistics chosen to hold value will subjectively determine the value of players. In the future, this model will be built upon this to determine what sort of players help a team win games to be more objective however there are a lot of things that cannot necessarily be found in the statistics.

Another ethical concern is that this program may affect the jobs of people in player development and management. This is a frequent concern in the world of machine learning so this program must used as a tool rather than a replacement in order to satisfy these needs.

Reflection

The project was an overall success. However, there were many lessons learned throughout the process, and questions that still are unanswered.

First of all, the results from the models were positive (considering the nature of the problem and the constraints on time), although it seems like it could be improved. The goaltender model performed very well, with a final validation accuracy of 87%, while the skater model modestly achieved a final validation accuracy of 31%[2]. This may be due to the fact that the statistics being analyzed in

[2] Note that the reported accuracies correspond to an error of within 5% of the prediction.

the project for goaltenders are much less variant in nature than for skaters. For instance, the save percentage statistic generally ranges from 0.9 to 0.95, while the number of points produced in one year for skaters can vary from 0 to 50. Other factors that could have impacted this include the effectiveness of how the data was inputted, and potentially not including enough features for the model to train on. Also, the goaltender model was prioritized first and less time was allocated for the skater model, which could be another reason as to why these results are much different. This is something that can be improved and analyzed in future endeavors.

Another obstacle encountered in this project was the handling of data. There was a heavy consultation with hockey statistics and analytics literature, although none gave a deep foundational knowledge of the correlation between statistics. For instance, is there a correlation between blocking shots and scoring goals? Data preprocessing was used to look for such relationships, by ignoring one statistic and evaluating the performance of the model. This was effective, however, visualizing the data would give more intuition onto why this occurred, and is something that could have been done in the process.

Finally, there was often trouble when considering how to input the data into the models. The model was trained on individual samples, without implementing a system of including player history, or including the players as a team. This could have drastically increased our results, especially for skaters. The processing of the data is something that can be improved in future projects.

References

1. McFarlane, Brian. *60 Years of Hockey: The Intimate Story Behind North America's Fastest, Most Exciting Sport : Complete Statistics and Records*. Toronto: McGraw-Hill Ryerson Ltd, 1976. Print.

2. McFarlane, Brian. *Everything You've Always Wanted to Know About Hockey: Histories of the National Hockey League and Stanley Cup*. New York: Scribner, 1971. Print.

3. Fraser, Don. *Power Play!: A Hockey Math Book*. Don Mills, Ont: Collier Macmillan Canada, 1978. Print.

4. McCurdy, Mica Blake. *Hockeyviz.com*

5. J. Brownlee, "When to Use MLP, CNN, and RNN Neural Networks," Machine Learning Mastery, 25-Apr-2018.