Data Mining Project Final Report NHL Playoff Prediction

Brendan Heal, Meara Kimball, Renee Fung, and Abdelrahman Alenazi

This report was written as part of the requirements for SENG 474 / CSC 578D: Data Mining, offered at the University of Victoria, taught by Dr. Alona Fyshe.

Abstract—What goes into winning the Stanley Cup in the National Hockey League (NHL)? The obvious things include deep offence, a smart defence and a goaltender who is able to make all the important saves. In this project we looked at quantifiable measures of some of those attributes and used algorithms to determine which team has the statistical advantage for the 2016 Stanley Cup Playoffs. We found an existing dataset and expanded on it using our own interpretation of the feature vectors. After creation of the combined dataset, we trained and tested several classifiers including an artificial neural network in single game prediction, and applied this to playoff series prediction.

1 Description and Motivation

DATA mining is already prevalent in many sports and has been shown to be extremely valuable to sports teams, fans and gambling organizations. This document outlines our groups plan to apply predictive analysis to the National Hockey League (NHL) in order to predict the winner of a particular game, based on statistics gathered over the season to that point. This includes the collection and processing of NHL data, classification using various algorithms, and finally application of a Neural Network to make a playoff prediction for the 2015-2016 NHL season.

Making accurate predictions using sports statistics is a very valuable application of data mining. It has value for teams looking to improve their performance, fans interested in the statistical side of the game or interested in sports gambling, and sports gambling businesses. Compared to other sports like soccer, baseball and basketball, statistical analysis in hockey is in its infancy. This is primarily to the fact that hockey is not as globally popular as other sports. But also, it can be difficult to analyze hockey as events such as goals and penalties occur so infrequently and games are often decided by luck.[1] There is potential for data mining to make a significant impact in the NHL.

2 Related Work

We have located two papers from University of Ottawa written by Joshua Weissbock, Herna Viktor and Diana Inkpen [2] and Joshua Weissbock alone [1]. Both papers used the same dataset collected by the first team over three months of the 2012/2013 NHL season. Both explored single game prediction. The more recent one also investigated prediction of playoff best-of-seven rounds. The first paper noted the dearth of previous hockey research in data mining, and recommended exploring the better researched area of soccer prediction, as it has similar statistics and event (goal) numbers. They reached a 59.38% success rate and later paper achieved 59.8% for single game predictions and 74% for the smaller set of playoff predictions.[2] They explored several methods and both found neural networks to generally be most successful.

We have contacted the team from the University of Ottawa and gained access to their original 2012/2013 dataset. We also gathered a second matching dataset using the current season from the hockey statistics website war-on-ice.com. Weissbock et al. [2] selected ten NHL statistics as features (see table). In order to predict the winner of a single hockey game, various cumulative statistics and performance statistics calculated from them are used as features. The cumulative statistics are collected from the beginning of the season until the date of the game which will be predicted. Both papers used two feature vectors per game, each containing the statistics for one team, with the result (win or loss) recorded for either the home or the away team. While we decided to use the same features as the original team, we stored both teams as a single feature vector for one game. This is further explained in Our Approach - Feature Vectors. The following statistics were selected:

Acronym	Defintion
GF	Goals For
GA	Goals Against
GD	Goal Differential
PP	Power Play Success Ratio
PK	Penalty Kill Success Ratio
SP	Shot Percentage
SvP	Save Percentage
WS	Win Streak
CS	Conference Standings
FC	Fenwick close
PDO	PDO
55GA	5v5 Goals For/Against

3 Our Approach

3.1 Feature Vectors

In the initial papers by Weissbock, Viktor and Inkpen, they used a single team's statistical snapshot as an instance of their data. This means that every individual game generates two separate feature vectors for the home team and the away team. The classifier must be run on both, and each team is predicted to win or lose the game. We found this problematic for several reasons.

First of all, there are no ties in an NHL hockey game, so these predictions must agree (predict win and lose respectively, or lose and win). The original researchers corrected this by measuring the strength of each prediction and running a script to check each game pair to check for ties. In the event of one, the stronger prediction would be kept and the weaker changed from win to lose or vice versa.

Another issue with the one game/two feature vectors method is the lack of attention to team interactions. Although teams aren't identified, we still believe there is information to be gained by comparing their statistics. As part of this, we also wanted to identify which team is home and away when competing against each other.

Our solution to all of these concerns was to combine the two teams' statistics into a single game instance. All the same features are present and in the same order, but we list first those of the home team and then those of the away team. Our classes must be likewise changed to match. Rather than win or lose, we now seek to determine if the game's winner will be the home team or the away team. Although home/away is not itself specified as a feature, the model will still be able to observe the home team advantage from this ordering. In general, we hoped this would allow the model to better capture the competition between two teams in their current state at the time of the game.

3.2 Dataset

3.2.1 Data Collection

As detailed above, our feature vectors will contain cumulative statistics for each team in a given hockey game. Therefore, NHL statistics must be arranged in the form of snapshots, for each team, directly before the game in question. Data in this form from the 2012/2013 season is provided by Joshua Weissbock, who worked on the research paper mentioned above. These snapshots are readily available for all seasons at war-on-ice.com/teambygame.html[3]. This team history page allows a user to plug in a number of metrics, including start and end dates, team, and statistic types. It then displays a table of statistics for each game in the date range, as well as cumulative statistics calculated over the date range. These tables can be downloaded as Comma Separated Value (CSV) files. Unfortunately, only a single team's cumulative statistics can be downloaded at a given time, and no underlying API was found which would allow the data to be queried using HTTP requests. In order to avoid manually collecting the data, desktop automation was used. The process of visiting this website, entering the appropriate dates and team for each game and then downloading the file was automated using a Python script. This script uses NHL schedule data, Selenium Web Driver and AutoIT in order to navigate the page and download the appropriate game information. Selenium Web Driver is used to drive the browser and interact with website elements, and AutoIT was used to provide keyboard input to a download confirmation dialog window in Firefox. The Python script then saves the data in a folder structure which is organized by date. The CSV files are named according to the game's date, team, and winner. They are additionally associated with a unique game id which was present in the original NHL schedule data. This allows the files to be re-associated with the NHL schedule data if additional information about the games is required.

3.2.2 Data Processing

In our data collection phase, we have collected a number of features that are more than what we need for the purpose of this project. Now the focus is manipulating the data to produce meaningful information and format it to the way that matches our desired feature vectors. A python script is implemented to extract selected features from files in pairs. Each file is named in the following format: year-month-day-gameID-Home Team-Away Team-[winner]-A or B.csv. Label A indicates that this data is for the home team, B indicates that this data is for the away team . eg. *example : 2015-10-08-11-COL-MIN-[COL]-B*. The script extracts the desired columns and produces feature vectors in a single file with a classifier in the final column.

3.3 Algorithm Selection

Selecting the appropriate algorithm for our predictive model is a critical step in achieving success in this project. It is known from past research that Neural Networks and Support Vector Machines typically show the best performance for sports predictions. This is generally due to the large amount of features in sports prediction models, and the complicated relationships between the features. We tested our data on a variety of classifiers provided by sklearn in Python. We also experimented with a Neural Network using a Netbeans based development environment called Neuroph Studio. We traned and tested our data using the following classifiers : Decision-Tree, Support Vector Machine (Linear Kernel), Naive Bayes (Gaussian), Stochastic Gradient Descent, and finally Neural Networks in Neuroph-Studio.

3.3.1 Playoff Predictions

Predicting is making claims about something that will happen based on information from the past and from the current state. There are two basic prediction type criteria: one, data that we have for teaching prediction and two, what we want to predict. There are several approaches for predictions and each has its own advantages and limitations. As mentioned above, our feature vectors were structured with the goal of predicting the winning team in mind. Each feature vector has a single classifier where 0 indicates that the home team wins, and 1 indicates that the away team wins.

In the playoffs, teams face off in seven game series, where the winner of the series continues to the next round. Since this project was due before the completion of the NHL season, we made a preliminary playoff prediction based on the conference standings as of March 15. The rounds of the playoffs are determined based on the standings, where the first place team plays the eighth place team, the second place team plays the seventh place team and so on. The higher placed team plays the first and last game of each series at home. Therefore, when applying our model to the playoffs, we predicted a single game where the higher placed team is considered to be the home team. In theory, the statistical nature of the game should improve our model's performance in the longer run of 7 games. This process was repeated for the subsequent rounds in order to determine a winner of the Stanley Cup.

3.3.2 Cross Validation

All algorithms were validated using k-Fold cross validation over 10 folds. This allows for much better validation of the models and a more reliable assessment of the performance of the model.

3.3.3 Neural Networks

Neural networks (NN) were inspired from brain modeling studies and can be used for prediction with various levels of success [4]. It is an adaptive system that learns to perform a function given an input and output map from a set of data. Unlike the dependencies in regression, the advantage of neural networks includes the automatic learning of dependencies only from measured data without any need to add further information.

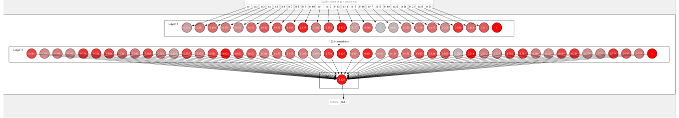
The objective of our problem is to create and train a neural network to predict the outcome of an NHL hockey game - which team is most likely to win - given some attributes as input. There are three main types of learning algorithms for training the neural network: supervised learning, unsupervised learning and reinforcement learning [5]. Supervised learning requires a training set and a desired output or target whereas unsupervised learning aims to find patterns in the input data with no assistance from an external source. Reinforcement learning, as the name suggests, rewards good performances and penalizes bad performances. For the purposes of this project, we used supervised learning as we have a training set of team statistics and an output classifier which represents the winning team.

3.3.4 NeurophStudio

NeurophStudio allows simple development of neural networks using a graphical IDE based on Netbeans.It allows a user to select from a variety of Neural network types, architechtures, transfer functions and learning rules to use. Following a report on a premier league prediction which was done using NeurophStudio[6], we made a Multi-Layer Perceptron using a Sigmoid transfer function and backpropogation with momentum as the learning rule.

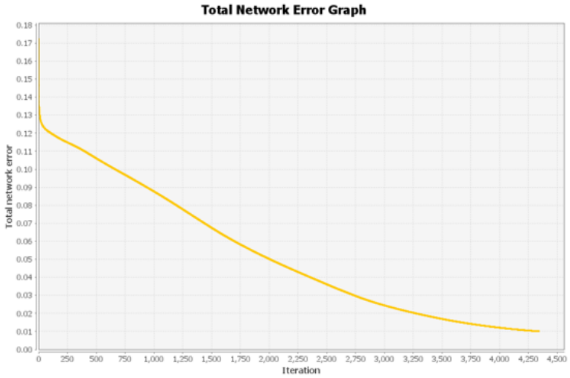
NeurophStudio allows a user to set a learning rate and momentum value to use when training the Neural Network. During training, NeurophStudio provides a graphical visualization of the root-mean squared error in the model at each training iteration. A successfully trained Neural Network needs to have a near zero error on the training set. After experimenting with various Neural Network architechtures and learning rates, we found the most successful Neural Network architecture to be one which consisted of a single hidden-layer and 48 hidden nodes.

Fig. 1. Multi-Layer Perceptron Architechture



Using this Neural Network architecture at a low learning rate of 0.01 and a momentum value of 0.7, the Neural Network reached an error value of 0.01 after 4310 iterations. This Neural Network ended up performing quite well when testing our data. However, NeurophStudio seems to be a somewhat unstable system for developing neural networks, and after some research, it has some well known performance issues which can cause the program to run very slowly and ultimately crash. This caused a lot of technical issues later on in the project for us. In the future, Encog seems to be a more suitable and reliable tool for developing Neural Networks.

Fig. 2. Training the Neural Network



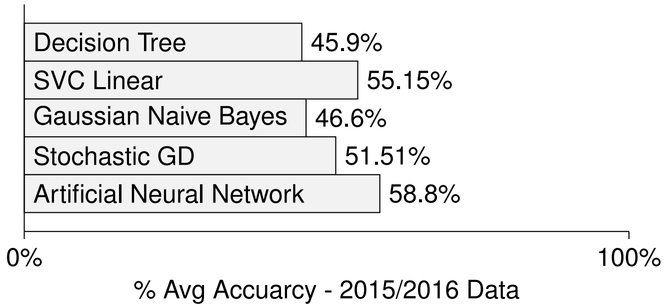
4 Results

What worked?

We ran our data on multiple algorithms such as Decision Tree, Linear Support Vector Machine SVC, Gaussian Naive Bayes, Stochastic Gradient Descent and Artificial Neural Network. This table shows the average accuracy and best case accuracy for each classifier over 10 fold cross-validation. These results are from the 2015/2016 season data we collected.

Classifier	Avg Accuracy (10 folds)	Best Accuracy
Decision Tree	45.9%	51.15%
SVC Linear	55.15%	62%
Gaussian NB	46.6%	50%
Stochastic GD	51.51%	62.12%
Artificial NN	58.8%	61.15%

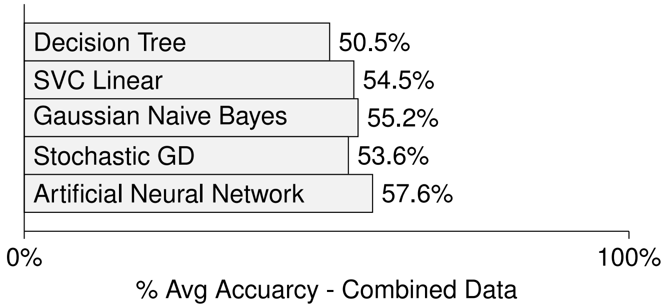
Fig. 3. 2015/2016 season dataset results



As you can see in Figure 3, Artificial Neural Network reached 58.8% which is within 1% of both orginal papers [1, 2]. However, you can see the variability in the results: for example ANN average was 58.8% where its best case accuracy was 61.15%. In this case, we responded to this variability issue we faced by adding more data for training and test classifiers again. Figure 4 shows better reliable results with more data were add (season 2013 and 2015/2016)

Classifier	Avg Accuracy (10 folds)	Best Accuracy
Decision Tree	50.5%	55%
SVC Linear	54.5%	58.9%
Gaussian NB	55.2%	58%
Stochastic GD	53.6%	58.9%
Artificial NN	57.6%	59.4%

Fig. 4. Combined data 2013 season & 2015/2016 season dataset results



What didn't work, and why?

Our final playoff prediction seemed to be questionable at best. Unfortunately, using an Artificial Neural Network to make the prediction makes it difficult to analyze what particular decision boundary lead the Network to predict in the way that it did. However, after thinking more about applying regular season statistics in order to make a playoff prediction, a number of potential problems in our approach can be seen. In the playoffs, teams play against the best teams in the league. However, in the regular season, teams play most of their games within their division. Some division have very strong teams, and other divisions have very weak teams. In some cases, a team's cumulative statistics may be artificially padded with wins and goals taken from weak teams that they play over and over again over the course of the season. In order to fix this potential shortcoming, an additional feature that takes into account a team's performance against stronger teams could be added. There are a number of other possibilities to be explored in order to strengthen the playoff prediction. However, we have not yet seen the result, and our prediction may turn out to be accurate.

Did you learn something new about the problem or about the algorithms you used?

We learned alot about Neural Networks during this project. Using an iterative process as well as some reasoning we were able to narrow down on an appropriate architecture for the Multi-Layer Perceptron. However, there is still allot of experimentation and further research to be done in determining all the various choices available when attempting to design a neural network. We learned that Artificial Neural Networks perform better with more data but it also comes with a computational trade off, the more train data you have for the ANN, the better machine you need to use to do this experiment. Because the Neural Network takes alot of computation power, and NeurophStudio seemed to be bugged, experimenting on various network architectures was slow and painful. More training data stabilized the results and reduce the variability.

If you created a dataset, did it turn out to be useful?

Yes we collected a 2015/2016 season data and clean by extracting most needed features. Also we have normalized the data using the min-max normalization technique.[7] 1

$$A' = (\frac{A - minValueOfA}{maxValueOfA - minValueOfA}) * (D - C) + C \tag{1}$$

where,
A′ contains Min-Max Normalized data one
if pre defined boundary is[C,D]
if A is the range of original data
& B is the mapped one data then,

Now our data is useful for classification purposes, as well as future NHL research.

Is there additional information which would have been useful?

There is tons of additional data which may have strengthened our model. Adding data from the 2013 season seemed to stabilize our results, so having even more seasons of data would be beneficial. As mentioned above, additional features may improve our playoff predictions, such as a teams performance against strong teams and individual player information. Player performance and events such as player injury are not captured in the model and may have a significant effect on the result of a game.

5 Conclusion and Future Work

Based on the NHL playoff format as of 2014, the Stanley Cup Playoffs includes 16 teams, eight from the West and eight from the East. The top three teams in each division are automatically qualified for the postseason. The remaining four teams will qualify through wild card spots which are given to teams with the most remaining points. Running our dataset through our neural network, we found the following matchups:

Round 1	**Team A**	**Team B**
West	Minnesota Wild	Dallas Stars
West	St. Louis Blues	San Jose Sharks
West	Los Angeles Kings	Nashville Predators
West	Anaheim Ducks	Chicago Blackhawks
East	Tampa Bay Lightning	Pittsburgh Penguins
East	New York Islanders	Florida Panthers
East	Washington Capitals	Philadelphia Flyers
East	New York Rangers	Boston Bruins
Round 2	**Team A**	**Team B**
West	Dallas Stars	San Jose Sharks
West	Los Angeles Kings	Chicago Blackhawks
East	Pittsburgh Penguins	Florida Panthers
East	Washington Capitals	Boston Bruins
Conf.	**Team A**	**Team B**
West	Dallas Stars	Los Angeles Kings
East	Florida Panthers	Washington Capitals
Final	**Team A**	**Team B**
Final	Dallas Stars	Florida Panthers

The two teams in the final predicted by our model were Dallas Stars for the Western Conference and the Florida Panthers for the Easter Conference with the Panthers winning the Stanley Cup for the 2015/2016 season.

There are quite a few things we can add to our list of future work to improve our prediction analysis for the NHL. Due to the closing deadline, our team was not able to collect all the team statistics for the 2015/2016 season as we set the cutoff date to March 15, 2016. This meant that each team had about eight to ten games left to play in the season when we ran our dataset through our model. Ideally, we would like to collect a full set of data for the season to make a more accruate predictions. We decided to not include player statistics in the current project state because we felt the data was too dynamic to capture. Players can get injured or traded throughout the season resulting in the fluctuation of those statistics. However, as part of our future work, we can investigate including this feature somehow as the players do make up the team and define how well they do in the season. We can also look into predicting the behaviour of individual players and their performance within their respective team as well as how well their team will play against others.

With more feature selections and analysis, we can utilize richer data with more sophisticated neural network configuration techniques to make more accurate predictions. Experimenting with more data, we plan to apply other classifiers and techniques for playoff predictions and compare them with each other for accuracy. For our dataset and classifier scripts, please visit https://github.com/brndnheal/SENG474

Like many sports, hockey is difficult to analyze given the random chance and luck that exists in every game but it makes the sport entertaining for fans to watch. We have learned the ins and outs of hockey and many other new facts about the role of statistics and chance in sports with an emphasis on how it is not always the best team who wins.

6 Project Timeline

Phase 1 - Preparation
Research, determine statistics and algorithm to use
Dataset gathering and processing
Progress: in final stages of dataset processing

Phase 2 - Model
Model implementation
Model training
Model testing
Progress: beginning implementation

Phase 3 - Report
Gather documentation
Give in-class presentation
Final report due April 6, 2016
Progress: gathering documentation

7 Team Organization

Phase 1:

All team members participated in research and organization. Brendan took the lead in data collection, Abdul with data processing and Renee in algorithm selection.

Phase 2:

The group reviewed the sklearn neural networks to begin modelling on the 2012/2013 dataset. All members will participate with implementation, either training or testing. Data processing was completed and data has been collected.

Phase 3:

Abdul and Brendan implemented a variety of classifiers in Sklearn and tested them with our data set. Renee and Meara researched the properties of the neural network and found information on NeurophStudio which was invaluable in putting together the neural network aspect of the presentation. Brendan trained the neural network, collected the final set of data to build the playoff prediction, and ran it through the neural network to make a playoff prediction. All group members helped put together the presentation and reports throughout the term, with Renee doing the design and layout for the presentation.

Overall, all group members contributed to the success of this project and worked seamlessly as a team.

References

[1] Joshua Weissbock, *Forecasting Success in the National Hockey League using In-Game Statistics and Textual Data*, Faculty of Engineering, University of Ottawa: 2014.

[2] Joshua Weissbock, Herna Viktor and Diana Inkpen, *Use of Performance Metrics to Forecast Success in the National Hockey League*, Faculty of Engineering, University of Ottawa: 2013.

[3] War-on-ice.com, *Team History*, www.war-on-ice.com/teambygame.html 2016. Accessed: January 2016

[4] Sandtander Meterology Group, *Data Mining and Artificial Intelligence: Bayesian and Neural Networks*, http://www.meteo.unican.es/research/datamining:2010. Accessed: February 2016

[5] Elnaz Davoodi and Ali Reza Khanteymoori, *Horse Racing Prediction Using Artificial Neural Networks*, Mathematics and Computer Science Department, Institute for Advanced Studies in Basic Sciences: 2010.

[6] neuroph.sourceforge.net,*Predicting the Result of Football Match with Neural Networks*, neuroph.sourceforge.net/tutorials/SportsPrediction/Premier League Prediction.html 2016. Accessed: March 2016

[7] S. Gopal Krishna Patro and Kishore Kumar sahoe, *Normalization: A Preprocessing Stage*, Department of CSE and IT, VSSUT, Burla, India : 2010.