**Predicting NHL match outcomes with ML models**

Gianni Pischedda

SoccerLogic Ltd

Uxbridge, London , UK

gianni.pischedda@soccerlogic.com

**Abstract**

Finding ways of predicting the outcome of sports games from performance data has always been an attractive proposition for many statisticians and, lately, of data miners using machine learning (ML) techniques. A research paper [1] on ice hockey (NHL) by a University of Ottawa team (from now on referred simply as Ottawa), and their generous sharing of the data used for their research provided the main drive for this paper.  
  
In this research, the Ottawa data is used for a number of purpose, all involving the use of ML techniques to predict the outcome of NHL games. First, we repeat Ottawa’s experiment, which looked “...at how effective traditional, advanced and mixed (both combined) statistics were for predicting success in the NHL”. Then we split all the given attributes in the data into *Categorical* and *Continuous*, and build ML separate models, whose result we compare with those of the original model. The original data is also parsed to create a new dataset, and models built to compare with the results of the original one. Lastly, a framework for making use of this data in a practical application (betting) is proposed, and the accuracy of models built is evaluate and compared. Three ML techniques: Decision Trees (DT), Artificial Neural Networks (ANN), and ClusteR, a software developed by a betting company, were used for these experiments.

**General terms**

NHL, Machine Learning, ML, Predictive modelling, *continuous* variables (Vars), *categorical* variables (Cat), data parsing. KnowledgeStudio, Angoss.

**Keywords**

NHL, Decision Trees, Bonferroni, Neural Networks, ClusteR, ML techniques, sabermetrics, k-nearest neighbour.

**1. Introduction**

Ice hockey (NHL) has not received much attention by sports analysts; not compared to other popular sports like football (NFL) or basketball (NBA), where much research has taken place over the years. A fact confirmed by Ottawa, who note that “As far as the authors are aware, there is no previous work, in the machine learning community, to predict the winner in a hockey game.” [1]

This is rather surprising given that ice hockey is a very popular sports in one of the most developed corners of the globe (Canada, US), and lots of data, including live match commentaries, is freely available on the NHL compared to other popular sports. This wealth of data was a major drive for this research.

A recent research to predict the outcome of Toronto matches during the same season (2012-2013), but using only match stats (Shots, Blocks, Saves, etc.), was an added incentive. Since good accuracies – over 70%, compared to Ottawa’s 60% - were achieved, the opportunity to try to improve on Ottawa’s results was hard to resists, especially given that no data gathering was required.

**2. Previous research**

Using ML to predict the outcome of a sport is not new; it has been done for a number of years, and with varying degree of success. Judging from the literature, neural networks appear to be the technique of choice. Chen et al. [2] were among the first to use them for picking winners in greyhound racing. And claim to have made a profit by betting on the indications of their model. Huang and Chang [3] also used neural network in soccer, to predict results of the 2006 World Cup, and claim of having reached accuracies of 75%. Neural networks were also used by Purucker [4] for predicting results in the NFL, and an accuracy of 78.6% was claimed. While Pardee [5] claims to have achieved 76% accuracy using neural networks to predicting NCAA football game results. He only studied teams in the Big Ten conference using the 1998 season statistics for training the algorithm, and the following season’s statistics for testing. As Chen et al [2] did for the greyhound racing study; Pardee chose to use back-propagation. His 76% beat the 68.2% prediction rate of a computer ranking system used by the BCS formula called the Massey Index Ratings.

Basketball, probably the sports which attracts most research, was studied by Loeffelholz et al. [5], who claim a 74.33% accuracy in predicting NBA games using neural networks on simple box score stats. More recently, Zimmerman et al [7] focused on basketball matches, for the NCAAB, and by comparing a number of ML techniques, they discovered the Naive Bayes provides the best accuracies. They also suggest that there is an upper-limit, a glass-ceiling of 74-75% in predicting results. This glass-ceiling seem to apply to predicting results of most sports, suing both ML and traditional statistical methods. In contrast, Ottawa [1] in what claims to be the first application of ML method to ice hockey (NHL) puts this figure at 60%. This 15% difference in accuracy between NHL and other sports does appear to provide plenty of room for improvement. And this was one of the aims of this research paper.

**3. A brief introduction to ICE HOCKEY**

In the Ottawa paper there is detailed background information regarding the game of ice hockey, and the data used. There is nothing of substance that can be added to this, as their description captures very well the relevant aspects of the game.

“In hockey, five players and a goalie per team are on an ice surface and play for a total of 60 minutes. The goal of the game is to put a rubber puck into the opposing team’s net using a 1.5 to 2m long stick made of wood or a composite material. The team who scores the most goals in a game is the winner. In the regular season, if a game is tied after 60 minutes, the teams play an extra 5 minutes of sudden death overtime and after that the game is decided by a shootout. In the playoffs, after 60 minutes, additional 20 minute overtime periods are played until a team scores.” [1]

**4. research objectives**

This research has the following main objectives:

* 1. To repeat Ottawa’s analysis and compare results
  2. To study the predictive powers of the two types of attributes in the Ottawa dataset: the categorical (Team, Location), and stats + metrics ones (all others). The accuracy of models built using these two, and that of both of them combined is evaluated and tested.
  3. To study the accuracy of the best performing model, as indicated by the previous tests, in a practical application, such as betting. In this popular pursuit, the aim is to predict (or to assign probabilities/odds) to the outcome of the next set of matches (one or two Days’ play) using current season previous Days’ play results – a Day-by-Days prediction.

**5. The INPUT data**

## 5.1 Datasets used

The dataset is the one used by Ottawa, and kindly shared. This is defined by V1 + V2 + Class in their paper, and VH,A in mine. The background of the data given by Ottawa, reads as follows: “The NHL is North America’s top hockey league comprising of 30 teams: 23 from the United States and 7 from Canada. Teams play a total of 82 games each during the regular season from October to April for a total of 1230 games... this data was collected for a period of nearly three months during the 2012-2013 NHL season, between 16 February and 28 April 2013, for a total of 517 games split over 72 Day’s play, and does not feature all the games (720) played during that season.”

This dataset consists of 1034 (517 matches) data points, each one showing cumulative stats and metrics by team updated after each game. A snapshot of the first few lines of this dataset is show in Table 1.

**Table 1. The first few lines of the Ottawa dataset**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Date | Team | Location | Gf | GA | GlDiff | PP% | PK% | Sh% | Sv% | Win Streak | Conf. Standing | Fenwick Close % | PDO | 5-5F/A | Class |
| 16/02/13 | Tampa Bay | Home | 49 | 40 | 9 | 26.4 | 81.1 | 881 | 920 | -1 | 9 | 42.99 | 1039 | 1.28 | Win |
| 16/02/13 | Florida | Away | 30 | 47 | -17 | 18.9 | 74.5 | 937 | 901 | -2 | 14 | 46.73 | 964 | 0.55 | Loss |
| 16/02/13 | Ottawa | Home | 35 | 27 | 8 | 17.6 | 89.3 | 932 | 945 | -1 | 7 | 53.19 | 1013 | 1.35 | Loss |
| 16/02/13 | Toronto | Away | 40 | 36 | 4 | 14.9 | 78.3 | 901 | 933 | -1 | 6 | 45.02 | 1032 | 1.22 | Win |
| 16/02/13 | Philadelphia | Home | 37 | 45 | -8 | 18.3 | 80.6 | 933 | 911 | -1 | 10 | 51.62 | 978 | 0.85 | Loss |
| 16/02/13 | Montreal | Away | 36 | 33 | 3 | 20.3 | 81.5 | 921 | 931 | 2 | 5 | 52.03 | 1010 | 1.11 | Win |

We can think of this dataset as made up by two data vectors: VH with the Home team stats, and VA for those of the Away team. From these two data vectors, a new dataset was derived, VH-A= VH-VA, where the stats now represent the difference between those of the Home and Away team. While the calculation for most of the variables was straightforward, this was not the case for Win Streak and Conf. Standing. Unlike the others, these variables are not quantitative values; Win Streak is an indicator of form and the difference gave the correct result. But since Conf. Standing represents relative strength or rank, calculating the difference gives the wrong results (negative when higher and positive when lower) and therefore the sign must be reversed to get the right value of this attribute.

The two datasets, VH,A (1034 – team stats) and VH-A (517 – match stats) are the ones that will form the basis of the datasets used for our analysis. The classifier (Class) Win/Loss was used for the first, and Home\_Win/ Away\_Win for the second.

.

**5.2 Attributes/variables**

A description of the variables used for each match is shown in Table 2.

**Table 2. Variables/attributes description**

|  |  |
| --- | --- |
| **Variable/attribute** | **Explanation** |
| Goals for | Total number of goals *scored* so far this season |
| Goals against | Total number of goals *conceded* so far this season |
| Goals Differential | Goals for – Goals against |
| Power Play Success Rate | Ratio – *scoring* a goal when 5 on 4 |
| Power Kill Success Rate | Ratio – not *conceding* a goal when 4 on 5 |
| Shot % | Goal *scored*/shots taken |
| Save % | Goals *conceded*/shots saved |
| Winning Streak | Number of consecutive games won |
| Conference Standing | Latest ranking on conference table |
| Fenwick Close % | Possession ratio |
| PDO | Luck parameter |
| 5/5 Goal For/Against | Ratio – 5 on 5 Goals For/Against |

It is unclear from the Ottawa paper [1] how these variables were *typed* - i.e. set to be of diverse types: categorical/continuous, ordered/unordered – if and when required by the ML techniques used. The Decision Trees (DT) technique used in this study requires *typing* input variables as either continuous or categorical, in order to make best use of them. Most variables in the Ottawa dataset are easily *typed* as except for Win Streak and Conf Standing. These do not indicate quantity, but order/rank, and therefore would normally be typed as ordered categorical or ordered continuous. Initially, both settings were tried, but when the results were the same, *ordered categorical* was chosen with default 10 equal-size intervals (20 were also tried, but with no difference in the results).

Excluding the classifier (Win/Lose), the *categorical* (Cats) variables in the data, are the Location (Home/Away) and Teams. All other variables were *typed* as *continuous* (Vars).

**6. ML techniques and software**

Two classical ML techniques were used for the analysis: Decision Trees (DT), and Multi-Layer Artificial Neural Networks (ANN), as implemented in the commercial software KnowledgeStudio (Angoss). Relevant parameter settings are specified as follows:

**DT**: CHAID, Exhaustive, No pruning, Bonferroni adjusted; All variables typed *Continuous*, with the exception of Team and Location (*Categorical, unordered*)

**ANN**: 8/10/12 Hidden neurons, Logistics Link Function, Conjugate gradient Optimizer, Grad Epsilon=0.0001, Line Epsilon: 0.9

A proprietary software, ClusteR, was also used. This belongs to a betting company that uses it for forecasting, setting odds, etc. for various sports. Details of this software cannot be disclosed for confidentiality reasons. All that can be revealed is that it combines k-nearest neighbour techniques with rule-finding ones.

**7. Model development and results**

## 6.1 Experiment 1

The aim of this experiment was to evaluate and compare the predictive accuracy of models constructed using the datasets and the ML techniques previously described. Models with both datasets, **VH,A** and V**H-A**, and all the variables (Cat+Vars) were created and evaluated.. These datasets are also split into two: one with only the *categorical variables* (Team and Location) and one with all others (GF,.., 5-5F/A), and named Cat and Vars. Models were also built and evaluated for these datasets. The reason behind this variable split was to evaluate which (Cat or Vars) make the greatest contribution to the accuracy of the models. Overall this experiment involved building eighteen different models, whose accuracies are shown in the table below in Table 3.

The train/test data partition sizes used in model building was approximately 65/35 percent, equivalent to Day’s play: 1-48 and 49-72 respectively. Ottawa used a 66/33 percent partition, and does not mention to have been concerned about splitting a Day’s play between the two

.

**Table 3. Models accuracies**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **DT** | | **ANN** | | **ClusteR** | |
| **Var type** | VH,A (1034) | VH-A (517) | VH,A (1034) | VH-A (517) | VH,A (1034) | VH-A (517) |
| **Cat** | 57.67 | 56.21 | 58.29 | 56.21 | 56.21 | 59.17 |
| **Vars** | 52.4 | 56.21 | 57.73 | 59.17 | 56.21 | 57.99 |
| **Cat+Vars** | 57.6 | 56.21 | 57.73 | 59.17 | 57.47 | 61.54 |

The results in Table 3 show that there is little difference between the accuracies obtained with the two datasets. But the VH-A dataset gives slightly better ones, 57.99 vs. 56.81. This dataset also provides the top accuracy when used with ClusteR (61.54%). Possibly the most interesting result to come out of this experiment is that models developed using only Cat variables match the accuracies of those of performance variables (Vars). In fact, Cat variables make the greatest contribution to model accuracies, as Location always is the most significant variable in all DT models – the only ones that provide clearly observable tree models. ANN, on the other hand, has the highest Vars accuracy, which suggest that is able to make the most of continuous variables

Table 4 gives a clearer picture of the results in Table 3. It shows the average of each column, and the top accuracy achieved

.

**Table 4. Average and best accuracy obtained (all models)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **DT** | | **ANN** | | **ClusteR** | |
| **Var type** | VH,A (1034) | VH-A (517) | VH,A (1034) | VH-A (517) | VH,A (1034) | VH-A (517) |
| **avg. All** | 55.89 | 56.21 | 57.92 | 58.18 | 56.21 | 59.57 |
| **Best** | 57.67 | 56.21 | 58.29 | 59.17 | 57.47 | 61.54 |

It is obvious from these results that the V**H-A** dataset provides the best accuracies independently of the technique used, and was chosen for the experiments that follow.

## 6.2 Comparison with Ottawa results

An indication of how the results of models built with the dataset VH,A compare with those of Ottawa using the same are given in Table 5. Note that this is the only dataset Ottawa used. As can be seen, there is little difference between the results.

**Table 5. Comparison with the V**H,A **dataset**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DT** | **ANN** | **SMO** |
| **Ottawa** | 55.42 | 57.06 | 58.61 |
| **Experiment 1** | 57.67 | 58.28 | 57.47 |
|  | **DT** | **ANN** | **ClusteR** |

There is, however, a more marked difference when Ottawa results are compared with those of the models built with the dataset V**H-A.**. Table 5. show that with this dataset better accuracy than Ottawa are achieved with both DT and ANN, with ClusteR showing the highest one. (Note that SMO has been compared to ClusteR solely because is Ottawa’s best performing technique.)

**Table 6. Comparison with VH-A dataset**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DT** | **ANN** | **SMO** |
| **Ottawa** | 55.42 | 57.06 | 58.61 |
| **Experiment 1** | 56.21 | 59.17 | 61.54 |
|  | **DT** | **ANN** | **ClusteR** |

The results of this experiment basically confirm Ottawa’s conjecture that there is a “glass-ceiling” of 60% accuracy in predicting NHL games: *“We found in our previous experiments that no matter what we tried we were not successful in predicting the NHL with accuracy higher than 60%.* [1].

## 6.3 Experiment 2 – Application to betting

In the preceding experiments, ML techniques have been used to build models for predicting match results in the NHL in a given season. The classical procedure of partitioning all the data into a test and train dataset has been used for creating and validating these models, and the most accurate ones highlighted.

This model construction, however, is not very useful in a practical context such as betting, whose aim is to predict the next set of matches at a particular date (Day’s play) from the results (performance data) of previous matches. For such purpose, the following procedure was created, and is outlined below:

1. Work out how many prior Days’ play data/stats are needed to build the first model.
2. Build models using different ML techniques
3. Use these models to predict the results of next Day’s play
4. Compare Actual vs. Forecast and evaluate models – choose best one with an accuracy > 62% (this is the upper bound specified in [3], and (assumed) minimum accuracy to bet profitably)
   1. If no result is better than 62%, add last Day’s results and go back to 2.
5. We now have a Betting Model
6. Update the data with the latest Day’s play results.
7. Start again from 2.

A similar one, which could act as an alternative or validation, could be used when close to the middle of the season, or enough data is available. Its objective would be to give more weight to recent results; a *form* model to compare performance with the *historic* modeloutlined by the above procedure.

While the data to build the *historic* model start from the first Day’s play (Day\_1), data for the *form* model would start after a few (K) Days play. So, for example, to forecast, say, the 17th Days play, data starting from Day K to 16 would be used. The choice of K is logically an important one, and likely to require the definition of some kind of heuristic. While this approach remains a direction for future research, in the following experiment K was chosen arbitrarily; logically with the only constraint that the resulting dataset would contain enough data points for a model to be built.

As we recently discovered, similar methods were employed by Kahn [9] for predicting the outcome of NFL games. Unfortunately this happened in the final review phase of the project, and there was no time to study it, compare it with ours, and draw lessons from it. It is comforting, however, to discover an historic precedent for this method.

The above theoretical construct is a reasonable one for the purpose, but one that does require a great deal of time and effort to implement. Time being at a premium, it has not been fully implemented in this experiment. In its place, a partial one was created. This consists of building a few *historic* and *form* models, to predict a number of matches at the beginning, middle at end of season, and study the resulting accuracies. These should give a good indication if the method can improve performance, and thus justify a major effort to implement it fully at a later date.

**Table 7. History model data and accuracies**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Experiment 2. - Data and accuracies | HISTORY models** | | | | | | | |
| Days play | Train Days | Train Matches | Test Day(s) | Test Matches | **DT** | **ANN** | **ClusteR** |
| 1-13 | 13 | 94 | 14-15 | 12 | 83.33 | 75.00 | 83.33 |
| 1-15 | 15 | 116 | 16-17 | 13 | 76.92 | 61.54 | 84.62 |
| 1-17 | 17 | 129 | 18-19 | 15 | 86.67 | 73.33 | 86.67 |
| 1-19 | 19 | 144 | 20-21 | 16 | 50.00 | 50.00 | 60.00 |
| 1-21 | 21 | 160 | 22 | 10 | 60.00 | 70.00 | 60.00 |
| 1-22 | 22 | 170 | 23 | 10 | 60.00 | 70.00 | 70.00 |
| 1-23 | 23 | 180 | 24-25 | 13 | 53.85 | 46.15 | 46.15 |
| 1-25 | 25 | 193 | 26-27 | 14 | 64.29 | 71.43 | 78.57 |
| 1-27 | 27 | 207 | 28-29 | 16 | 50.00 | 75.00 | 68.75 |
| 1-29 | 29 | 223 | 30-31 | 11 | 81.82 | 72.73 | 81.82 |
| 1-31 | 31 | 234 | 32-33 | 14 | 57.14 | 50.00 | 71.43 |
| - | - | - | - | - | **65.82** | **65.02** | **71.91** |
| 1-41 | 41 | 288 | 42-43 | 15 | 66.67 | 40.00 | 66.67 |
| 1-43 | 43 | 303 | 44-45 | 14 | 57.14 | 57.14 | 64.29 |
| 1-45 | 45 | 317 | 46-47 | 12 | 66.67 | 50.00 | 66.67 |
| 1-47 | 47 | 329 | 48-49 | 16 | 62.50 | 62.50 | 62.50 |
| 1-49 | 49 | 345 | 50-51 | 18 | 61.11 | 72.22 | 70.59 |
| 1-51 | 51 | 363 | 52-53 | 14 | 64.29 | 64.29 | 64.29 |
| - | - | - | - | - | **63.06** | **57.69** | **65.84** |
| 1-61 | 61 | 438 | 62-63 | 15 | 73.33 | 73.33 | 73.33 |
| 1-63 | 63 | 453 | 64-65 | 16 | 62.50 | 43.75 | 62.50 |
| 1-65 | 65 | 469 | 66-67 | 15 | 73.33 | 53.33 | 73.33 |
| 1-67 | 67 | 484 | 68-69 | 15 | 53.33 | 73.33 | 66.67 |
| 1-69 | 69 | 499 | 70-72 | 18 | 50.00 | 55.56 | 50.00 |
|  |  |  |  |  | **62.50** | **59.86** | **65.17** |
|  |  |  |  | **Avg** | **64.31** | **61.85** | **68.74** |
|  |  |  |  | **Min** | 50.00 | 40.00 | 46.15 |
|  |  |  |  | **Max** | 86.67 | 75.00 | 86.67 |

Table 7 shows details of the history models built and their accuracies. Average accuracies are higher than achieved in Experiment 1, 65% vs. 59%, and ClusteR gives the best overall results (avg. 68.74). Surprisingly, ANN fares worse than DT (61.85 vs. 64.31), and also show a significant drop in performance from early season matches, to the middle and end. This is probably explained by the facts that, in contrast with the other two techniques, ANN are very sensitive to parameter settings. Given the large numbers of models to be evaluated, time constraints did not allow optimisation to be performed. Surprisingly, all techniques perform significantly better early in the season than in middle and at the end. This general pattern is contrary to the orthodox view that more data equal better model.

**Table 8. Form model data and accuracies**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Experiment 2. - Data and accuracies | FORM models** | | | | | | | |
| Days play | Train Days | Train Matches | Test Day(s) | Test Matches | **DT** | **ANN** | **ClusteR** |
| 6-19 | 13 | 102 | 16 | 20-21 | 50.00 | 50.00 | 50.00 |
| 8-21 | 13 | 105 | 10 | 22 | 60.00 | 70.00 | 70.00 |
| 10-22 | 12 | 95 | 10 | 23 | 60.00 | 60.00 | 60.00 |
| 11-23 | 12 | 100 | 13 | 24-25 | 53.85 | 46.15 | 61.54 |
| 13-25 | 12 | 100 | 14 | 26-27 | 64.29 | 50.00 | 64.29 |
| 15-27 | 12 | 100 | 16 | 28-29 | 56.25 | 68.75 | 75.00 |
| 16-29 | 13 | 107 | 11 | 30-31 | 81.82 | 63.64 | 90.91 |
| 18-31 | 13 | 105 | 14 | 32-33 | 57.14 | 42.86 | 57.14 |
| - | - | - | - | - | **60.42** | **56.43** | **66.11** |
| 28-41 | 13 | 100 | 15 | 42-43 | 66.67 | 57.14 | 60.00 |
| 30-43 | 13 | 99 | 14 | 44-45 | 57.14 | 58.33 | 57.14 |
| 32-45 | 13 | 102 | 12 | 46-47 | 66.67 | 75.00 | 66.67 |
| 34-47 | 13 | 100 | 16 | 48-49 | 62.50 | 72.22 | 62.50 |
| 36-49 | 13 | 103 | 18 | 50-51 | 61.11 | 72.22 | 72.22 |
| 38-51 | 13 | 105 | 14 | 52-53 | 64.29 | 71.43 | 71.43 |
| - | - | - | - | - | **63.06** | **67.72** | **64.99** |
| 48-61 | 13 | 100 | 15 | 62-63 | 73.33 | 66.67 | 66.67 |
| 49-63 | 13 | 105 | 16 | 64-65 | 62.50 | 62.50 | 56.25 |
| 51-65 | 13 | 105 | 15 | 66-67 | 73.33 | 80.00 | 80.00 |
| 53-67 | 13 | 106 | 15 | 68-69 | 53.33 | 60.00 | 60.00 |
| 55-69 | 13 | 108 | 18 | 70-72 | 50.00 | 55.56 | 61.11 |
| - | - | - | - | - | **62.50** | **64.95** | **64.81** |
|  |  |  |  | **Avg** | **61.80** | **62.24** | **65.41** |
|  |  |  |  | **Min** | 50.00 | 42.86 | 50.00 |
|  |  |  |  | **Max** | 81.82 | 80.00 | 90.91 |

Table 8. above shows details of the *form* models built and their accuracies. Average accuracies are higher than those achieved in Experiment 1, and ClusteR continues to be the he top performer, with 65.41. Also the performance hierarchy between the methods is the same, with ANN second, and DT the third. There are also some significant differences between model performances in the three periods studied, with each technique displaying a unique accuracy.

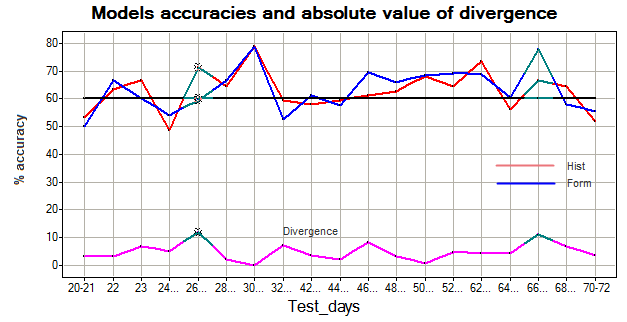
DT’s accuracy is pretty much constant, with only a 2% difference between periods. ANN, instead, rises sharply in middle season, and then loses some gains in the end: 56.43, 67.72, and 64.95. The small gap between DT and ANN accuracies, and with Location continuing to be the most significant variable, suggests that Vars (performance variables) make a small contribution to accuracy. ClusteR seems to be the exception to this, whereas one would have expected ANN to perform best.

**Table 9. History and form models compared**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **DT** | | **ANN** | | **ClusteR** | |
| Test Day(s) | Test Matches | **History** | Form | **History** | Form | **History** | Form |
|  |  | 83.33 |  | 75.00 |  | 83.00 |  |
|  |  | 76.92 |  | 61.54 |  | 84.62 |  |
|  |  | 86.67 |  | 73.33 |  | 86.67 |  |
| 20-21 | 16 | 50.00 | 50.00 | 50.00 | 50.00 | 60.00 | 50.00 |
| 22 | 10 | 60.00 | 60.00 | 70.00 | 70.00 | 60.00 | 70.00 |
| 23 | 10 | 60.00 | 60.00 | 70.00 | 60.00 | 70.00 | 60.00 |
| 24-25 | 13 | 53.85 | 53.85 | 46.15 | 46.15 | 46.15 | 61.54 |
| 26-27 | 14 | 64.29 | 64.29 | 71.43 | 50.00 | 78.57 | 64.29 |
| 28-29 | 16 | 50.00 | 56.25 | 75.00 | 68.75 | 68.75 | 75.00 |
| 30-31 | 11 | 81.82 | 81.82 | 72.73 | 63.64 | 81.82 | 90.91 |
| 32-33 | 14 | 57.14 | 57.14 | 50.00 | 42.86 | 71.43 | 57.14 |
| - | - | **59.64** | **60.42** | **63.16** | **56.43** | **67.09** | **66.11** |
| 42-43 | 15 | 66.67 | 66.67 | 40.00 | 57.14 | 66.67 | 60.00 |
| 44-45 | 14 | 57.14 | 57.14 | 57.14 | 58.33 | 64.29 | 57.14 |
| 46-47 | 12 | 66.67 | 66.67 | 50.00 | 75.00 | 66.67 | 66.67 |
| 48-49 | 16 | 62.50 | 62.50 | 62.50 | 72.22 | 62.50 | 62.50 |
| 50-51 | 18 | 61.11 | 61.11 | 72.22 | 72.22 | 70.59 | 72.22 |
| 52-53 | 14 | 64.29 | 64.29 | 64.29 | 71.43 | 64.29 | 71.43 |
| - | - | **63.06** | **63.06** | **57.69** | **67.72** | **65.84** | **64.99** |
| 62-63 | 15 | 73.33 | 73.33 | 73.33 | 66.67 | 73.33 | 66.67 |
| 64-65 | 16 | 62.50 | 62.50 | 43.75 | 62.50 | 62.50 | 56.25 |
| 66-67 | 15 | 73.33 | 73.33 | 53.33 | 80.00 | 73.33 | 80.00 |
| 68-69 | 15 | 53.33 | 53.33 | 73.33 | 60.00 | 66.67 | 60.00 |
| 70-72 | 18 | 50.00 | 50.00 | 55.56 | 55.56 | 50.00 | 61.11 |
| - | - | **62.50** | **62.50** | **59.86** | **64.95** | **65.17** | **64.81** |
|  | **Avg** | **61.47** | **61.8** | **60.57** | **62.24** | **66.19** | **65.41** |
|  | **Min** | 50.00 | 50.00 | 40.00 | 42.86 | 46.15 | 50.00 |
|  | **Max** | 81.82 | 81.82 | 75.00 | 80.00 | 81.82 | 90.91 |

A comparison between the two models is given in Table 9. and shows that there is only a small difference ( < 2%) in overall averages. Days’ model predictions, as well as differences between them, are plotted in Figure. 1, as are their differences, which reach a maximum of 11% (Fig.1). The models performance in the three periods examined shows some differences worth pointing out. DT starts lower, rises for middle season matches, and then stays pretty much the same at the end. ClusteR, instead, starts highest, and ends up lower. In both cases, however, the differences are minimal. In contrast, ANN shows some large and contrasting differences between the models in the three periods. History starts highest, then falls down, and then rises again. This pattern is reversed for the form model. The result is that ANN has the highest gap in accuracies between the three periods: 7, 10, and 5%. A likely explanation of this behaviour has already been mentioned.

**Figure 1. Accuracies (average of all techniques) of History and Form models compared, and their difference**

****

**7. Conclusions**

Let’s recall the aims of this research. The first was to compare models accuracies with Ottawa when using the same dataset and techniques. Slightly higher accuracies were achieved by this research, which basically confirms their results. Even higher accuracies, however, were achieved by all techniques when using the data parsed in a different way (subtracting values), and the 60% glass-ceiling breached with the 61.54% of ClusterR. The higher accuracies achieved with the VH-A dataset seems to suggest experimenting with different dataset constructs for achieving higher accuracies.

The second was to test the predictive accuracy of the different types of variables present in the Ottawa dataset. The results (Tables 3. and 4.) show that simple categorical variables like Team and Location drive much of the prediction accuracy of the models, whatever the technique used – complex metrics and stats on performance appear not to be such good predictors of success as it is normally assumed. The lesson here is not to underrate the importance of context variables in building such models.

Defining and testing a real-world application (betting) of ML classifiers was the third aim of this research. Here the objective was to build many models in sequence for predicting the outcome of the next round (Day) of matches in a running sport competition, be it hockey, football or basketball. A dynamic model building method, Day-by-Days prediction, was therefore invented and outlined. This method takes into account the practical aspects of betting – that of having to predict the next Day’s play results with the current season previous ones. Because of time constraints, it was not possible to fully test this method, but the results achieved - significantly better than those of the first experiments – seems to suggest that the effort required to do a complete test may be justified.

This project was also an opportunity to experiment with a proprietary classifications software: ClusteR, which combines k-nearest neighbour techniques with rule-finding ones, and compare it with traditional and established ones, such as DT and ANN. ClusteR outperformed both in our experiments, and did significantly better when dealing with small datasets (100 or less data points). Unsurprisingly, since it has been developed by a betting company, ClusteR is particularly useful in modelling a Day-by-Days prediction situation; when one may need to start build a good model with the smallest number of data points (Days’ play/matches) possible.

Achieving consistent accuracy was a problem in the third experiment. The randomness’ of the playing calendar may be the culprit here - not all teams play in the same Day, as it is normal in soccer leagues. Since the choice of Days Training and Days Testing was done somewhat randomly, the Train dataset may result to contain more/fewer matches than others Days’ play chosen. This selection may result in some matches being easier/harder to predict than others, and thus negatively affect accuracies. I wonder whether more consistent results may be achieved with a more realistic selection of Train data – one that takes into account the teams playing on the Day(s) to be predicted.

**8. A final reflection**

In some of the referenced papers, variable/feature selection in building models for predicting the outcome of sports is

much influenced by *sabermetrics* analysts. These are professionals with a deep knowledge of a particular sport but a scant one of ML techniques. Their focus on creating predictive complex metrics from basic performance stats appears somewhat in contrast with ML philosophy. It is not clear whether ML techniques learn better from a lot of simple stats, than from a few human pre-digested complex ones (metrics). Perhaps better results could be obtained if ML researchers used more primary data (stats + context: Teams, Location, etc.) in their experiments. One reason why *sabermetric* data is often preferred may be due to data collection weighing heavily onto ML researchers. Therefore it is far easier, as others and I have done, to avoid this tedious task, and use readily available (*saber*)metrics – of which plenty are readily available for many sports - to get quickly to the interesting one of building and testing models.

**9. ACKNOWLEDGMENTS**

The help of Albrecht Zimmermann in reviewing this research is gratefully acknowledged.

# 10. References

[1] Weissbock, J., Viktor, H., Inkpen, D.: Use of  
performance metrics to forecast success in the national hockey league. European Conference on Machine Learning: Sports Analytics and Machine Learning Workshop (2013)

[2] Chen, H., Buntin Rinde, P., She, L., Sutjahjo, S., Sommer, C., Neely, D.: Expert prediction, symbolic learning, and neural networks. An experiment on greyhound racing. IEEE Expert 9(6) (1994) 21–27

[3] Huang, K.Y., Chang, W.L.: A neural network method for prediction of 2006 World Cup football game. In: Neural Networks (IJCNN), IEEE (2010) 1–8

[4] Purucker, M.C.: Neural network quarterbacking. Potentials, IEEE 15(3) (1996) 9–15  
[5] Pardee, M.: An artificial neural network approach to college football prediction and ranking. University of Wisconsin (1999)

[6] Loeffelholz, B., Bednar, E., Bauer, K.W.: Predicting NBA games using neural networks. Journal of Quantitative Analysis in Sports 5(1) (2009) 1–15

[7] Albrecht Zimmermann, Sruthi Moorthy, Zifan Shi: Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned. European Conference on Machine Learning: Sports Analytics and Machine Learning Workshop (2013)

[8] Josh Weissbock, Diana Inkpen: Combining Textual Pre-game Reports and Statistical Data for Predicting Success in the National Hockey League. Canadian Conference on AI 2014: 251-262

[8] Kahn, Joshua, “Neural Network Prediction of NFL Games”, University of Wisconsin – Electrical and Computer Engineering Department, 2003.