**Model Trees for Identifying Exceptional Players in the NHL Draft**

Paper Track (Cambria, 14pt)

Paper ID

**Abstract**

Recruiting strong players is crucial for a team’s success. We describe a new data-driven interpretable approach for assessing draft prospects in the National Hockey League. Successful previous approaches have 1) built a *predictive model* based on player features [1], or 2) derived performance predictions from the observed performance of *comparable players* in a cohort [2]. This paper develops model tree learning, which incorporates strengths of both model-based and cohort-based approaches. A model tree partitions the feature space according to the values of discrete features, or learned thresholds for continuous features. Each leaf node in the tree defines a group of players, easily described to hockey experts, with its own group regression model. Compared to a single model, the model tree forms an ensemble that increases predictive power. Compared to cohort-based approaches, the groups of comparables are discovered from the data, without requiring a similarity metric. The performance predictions of the model tree are competitive with the state-of-the-art method. We show in case studies that the model tree player ranking can be used to highlight strong and weak points of players.

**1. Introduction**

Player ranking is one of the most studied subjects in sports analytics [17]. In this paper we consider predicting success in the National Hockey League (NHL) from junior league data, with the goal of supporting draft decisions. The publicly available junior league data aggregate a season’s performance into a single set of numbers for each player. Our method can be applied to any data of this type, for example also to basketball NBA draft data [27]. Since our goal is to support draft decisions by teams, we ensure that the results of our data analysis method can be easily explained to and interpreted by sports experts.

Previous approaches for analyzing hockey draft data take a regression approach or a similarity-based approach. Regression approaches build a predictive model that takes as input a set of player features, such as demographics (age, height, weight) and junior league performance metrics (goals scored, plus-minus), and output a predicted success metric (e.g., number of games played in the professional league). The current state-of-the-art is a generalized additive model [1]. Cohort-based approaches divide players into groups of comparables and predict future success based on a player’s cohort. For example, the PCS model [2] clusters players according to age, height, and scoring rates. One advantage of the cohort model is that predictions can be explained by reference to similar known players, which many domain experts find intuitive. For this reason, several commercial sports analytics systems, such as Sony’s Hawk-Eye system, identify groups of comparables for each player. Our aim in this paper is to describe a new model for draft data that achieves the best of both approaches, regression-based and similarity-based.

Our method uses a *model tree* [26,LogiBoost reference]. Each node in the tree defines a new yes/no question, until a leaf is reached. Depending on the answers to the questions, each player is assigned a group corresponding to a leaf. The tree builds a different regression model for each leaf node. The model tree offers several advantages.

1. Compared to a single regression model, the tree learns an *ensemble of regression models*, based on non-linear thresholds. This increases the expressive power and predictive accuracy of the model. The tree can represent complex interactions between player features and player groups. For example, if the data indicate that players from different junior leagues are sufficiently different to warrant building distinct models, the tree can introduce a split to distinguish different leagues.
2. Compare to a similarity-based model, tree construction *learns groups of players* from the data, without requiring the analyst to specify a similarity metric. The distinctions between the groups are guaranteed to be predictively relevant to future NHL success. Also, the tree creates a model, not a single prediction, for each group, which allows it to differentiate among players from the same group.

A natural approach would be to build a linear regression tree to predict NHL success, which could be measured by the number of games a draft pick plays in the NHL. However, only about half the draft picks ever play a game in the NHL [19]. As observed by [1], this creates a zero-inflation problem that limits the predictive power of linear regression. We propose a novel solution: instead of predicting the number of games played, we use logistic regression to predict whether a player will *play at least one game in the NHL*. We learn a logistic regression model tree for the NHL draft data, and rank players by the probability that the logistic regression model tree assigns to them playing at least one game. Intuitively, if the more confident we are that a player will play at least one NHL game, the more games we can expect him the play.

Following [1], we evaluate the logistic regression ranking by comparing it to ranking players by their future success, measured as the number of NHL games they play after 7 years. The correlation of the logistic regression ranking with future success is competitive with that achieved by the generalized additive model of [1]. We show in case studies that the logistic model tree adds information to the NHL’s Central Scouting Service Rank (CSS). For example, Stanley Cup winner Kyle Cumiskey was not ranked by the CSS in his draft year, but was ranked as the third draft prospect in his group by the model tree, just behind Brad Marchand and Mathieu Carle. Our case studies also show that the feature weights learned from the data can be used to explain the ranking in terms of which player features contribute the most to an above-average ranking. In this way the model tree can be used to highlight exceptional features of a player for scouts and teams to take into account in their evaluation.

*Paper Outline.* After we review related work, we show and discuss the model tree learned from the 2004-2006 draft data. The rank correlations are reported to evaluate predictive accuracy. We discuss in detail how the ensemble of group models represents a rich set of interactions between player features, player categories, and NHL success. Case studies give examples of strong players in different groups and show how the model can used to highlight exceptional player features.

**2. Previous Work**

Player ranking is one of the most studied subjects in sports analytics [17]. Different approaches are appropriate for different data types. For example, with dynamic play-by-play data, Markov models have been used to rank players [12, 13, 14, 18]. For data that record the presence of players when a goal is scored, regression models have also been applied to extend the classic plus-minus metric [15, 16]. In this paper, we utilize player statistics that aggregate a season’s performance into a single set of numbers. While this data is much less informative than play-by-play data, it is easier to obtain, interpret, and process. Our goal is to predict NHL performance from junior league data, for which only aggregate statistics are publicly available. A different problem is predicting future NHL performance from past NHL performance [3].

*Regression Approaches*. To our knowledge, this is the first application of model trees to hockey draft prediction. The closest predecessor to our work is the generalized additive model due to Shuckers [1] for predicting future NHL game counts from junior league data. The ranking that results from the additive model shows a higher correlation with the ranking from actual NHL game counts, than the order in which players were drafted. To our knowledge, the model of this paper is the first to predict whether a draftee plays any games at all, rather than predicting his total game count.

*Similarity-Based Approaches* use clustering to group similar players and predict performance. In the Prospect Cohort Success (PCS) model [2], cohorts of draftees are defined based on age, height, and scoring rates. Model tree learning provides an automatic method for identifying cohorts with predictive validity. We refer to cohorts as groups to avoid confusion with the PCS concept. Because tree learning is computationally efficient, our model tree is able to take into account a larger set of features than age, height, and scoring rates. Also, it provides a separate predictive model for each group that assigns group-specific weights to different features. In contrast, PCS makes the same prediction for all players in the same cohort. So far, PCS has been applied to predict where a player will score more than 200 games career total. Tree learning can easily be modified to make predictions for any game count threshold.

*Model-based clustering* is a fundamental statistical technique for identifying player groups in such a way that a predictively accurate model can be built for each group [23, 24]. Model-based clustering algorithms assign players to arbitrary groups, which are often difficult to interpret in terms of player features. Because clusters are not defined by player features, the clustering does not support making predictions for new players that have not been assigned to a cluster, unlike a model tree.

**3. Dataset**

Our data was obtained from public-domain on-line sources, including nhl.com, eliteprospects.com, and draftanalyst.com. We are also indebted to David Wilson for sharing his NHL performance dataset [3]. The full dataset is posted on the worldwide web [4]. We consider players drafted into the NHL between 1998 to 2008. Following [1], we took as our dependent variable **the total number of games *gi*** played by a player *i* after 7 years under an NHL contract. The first seven seasons are chosen because NHL teams have at least seven-year rights to players after they are drafted [9]. Our dataset includes also the total time on ice after 7 years. The results for time on ice were very similar to number of games, so we discuss only the results for number of games. The independent variables include demographic factors (e.g. age), performance metrics for the year in which a player was drafted (e.g., goals scored), and the rank assigned to a player by the NHL Central Scouting Service (CSS). If a player was not ranked by the CSS, we assigned (1+ the maximum rank for his draft year) to his CSS rank value. Another preprocessing step was to pool all European countries into a single category. Table 1 lists all data columns and their meaning. Table 2 shows an excerpt from the dataset.

|  |  |
| --- | --- |
| Variable Name | Description |
| id | nhl.com id for NHL players, otherwise Eliteprospects.com id |
| DraftAge | Age in Draft Year |
| Country | Nationality. Canada -> 'CAN', USA -> 'USA', countries in Europe -> 'EURO' |
| Position | Position in Draft Year. Left Wing -> 'L', Right Wing -> 'R', Center -> 'C', Defencemen -> 'D' |
| Overall | Overall pick in NHL Entry Draft |
| CSS\_rank | Central scouting service ranking in Draft Year |
| rs\_GP | Games played in regular seasons in Draft Year |
| rs\_G | Goals in regular seasons in Draft Year |
| rs\_A | Assists in regular seasons in Draft Year |
| rs\_P | Points in regular seasons in Draft Year |
| rs\_PIM | Penality\_in\_Minutes in regular seasons in Draft Year |
| rs\_PlusMinus | Goal Differential in regular seasons in Draft Year |
| po\_GP | Games played in playoffs in Draft Year |
| po\_G | Goals in playoffs in Draft Year |
| po\_A | Assists in playoffs in Draft Year |
| po\_P | Points in playoffs in Draft Year |
| po\_PIM | Penalty\_in\_Minutes in playoffs in Draft Year |
| po\_PlusMinus | Goal differential in playoffs in Draft Year |
| sum\_7yr\_GP | Total NHL games played in player's first 7 years of NHL career |
| sum\_7yr\_TOI | Total NHL Time\_on\_Ice in player's first 7 years of NHL career |
| GP\_7yr\_greater\_than\_0 | Played a game or not in player's first 7 years of NHL career |

Table 1 Player Attributes listed in dataset.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | Player  Name | Draft  Age | Coun  try | Hei  ght | Wei  ght | Posi  tion | Over  all | CSS\_  rank | rs  \_GP | rs\_  G | rs\_  A | rs\_  P | rs\_  PIM | rs\_  Plus  Minus | sum\_  7yr\_  GP | sum\_  7yr\_  TOI | GP\_  7yr > 0 |
| 846-  9455 | Jason  Speeza | 18 | CAN | 75 | 210 | C | 2 | 1 | 63 | 46 | 76 | 122 | 45 | 0 | 322 | 5714 | yes |
| 846-  9474 | Colby  Armstrong | 19 | CAN | 74 | 185 | R | 21 | 8 | 76 | 36 | 43 | 79 | 168 | 0 | 199 | 3397 | yes |
| 846-  9454 | Ilya Kovalchuk | 18 | EURO | 75 | 230 | R | 1 | 1 | 53 | 43 | 24 | 67 | 141 | 0 | 466 | 9905 | yes |
| 847-  0105 | Scottie Upshall | 19 | CAN | 72 | 200 | R | 6 | 5 | 68 | 35 | 54 | 89 | 149 | 0 | 230 | 2998 | yes |
| 847-  1675 | Sidney Crosby | 18 | CAN | 71 | 200 | C | 1 | 1 | 73 | 78 | 110 | 188 | 94 | 0 | 434 | 9144 | yes |
| 847-  4141 | Patrick Kane | 19 | USA | 71 | 177 | R | 1 | 2 | 65 | 67 | 87 | 154 | 94 | 44 | 515 | 9927 | yes |
| 847-  3419 | Brad Marchand | 18 | CAN | 69 | 181 | L | 71 | 80 | 68 | 29 | 37 | 66 | 83 | 40 | 218 | 3418 | yes |
| 27 | Yared Hagos | 18 | EURO | 73 | 218 | C | 70 | 24 | 43 | 11 | 26 | 37 | 24 | 1 | 0 | 0 | no |

Table 2 Sample Player Data. rs = regular season. We use the same statistics for the playoffs (not shown).

**4. Model Tree Construction**

Model trees are a flexible formalism that can be built for any regression model. An obvious candidate for a regression model would be linear regression; alternatives include a generalized additive model [1], and a Poisson regression model specially built for predicting counts [25]. We introduce a different approach: *a logistic regression model to predict whether a player will play any games at all in the NHL (gi>0).* The motivation is that many players in the draft never play any NHL games at all (up to 50% depending on the draft year) [19]. This poses an extreme zero-inflation problem for any regression model that aims to predict directly the number of games played. In contrast, for the *classification problem* of predicting whether a player will play any NHL games, zero-inflation means that the data set is balanced between the classes. This classification problem is interesting in itself; for instance, a player agent would be keen to know what chances their client has to participate in the NHL. The logistic regression probabilities *pi=P(gi>0)* can be used not only to predict whether a player will play any NHL games, but also to *rank players* such that the ranking correlates well with the actual number of games played. Our method is therefore summarized as follows.

1. Build a tree whose leaves contain a logistic regression model.

2. The tree assigns each player *i* to a unique leaf node *li*, with a logistic regression model *m(li)*.

3. Use *m(li)* to compute a probability *pi= P(gi>0)*.

Figure 1 shows the logistic regression model tree learned for our second cohort by the LogiBoost algorithm (described in the appendix). It places CSS rank at the root as the most important attribute. Players ranked better than 12 form an elite group, of whom almost 82% play at least one NHL games. For players at rank 12 or below, the tree considers next their regular season points total. Players with rank and total points below 12 form an unpromising group: only 16% of them play an NHL game. Players with rank below 12 but whose points total is 12 or higher, are divided by the tree into three groups according to whether their regular season plus-minus score is positive, negative, or 0. (A three-way split is represented by two binary splits). If the plus-minus score is negative, the prospects of playing an NHL game are fairly low at about 37%. For a neutral plus-minus score, this increases to 61%. For players with a positive plus-minus score, the tree uses the number of playoff assists as the next most important attribute. Players with a positive plus-minus score and more than 10 playoff assists form a small but strong group that is 92% likely to play at least one NHL game.

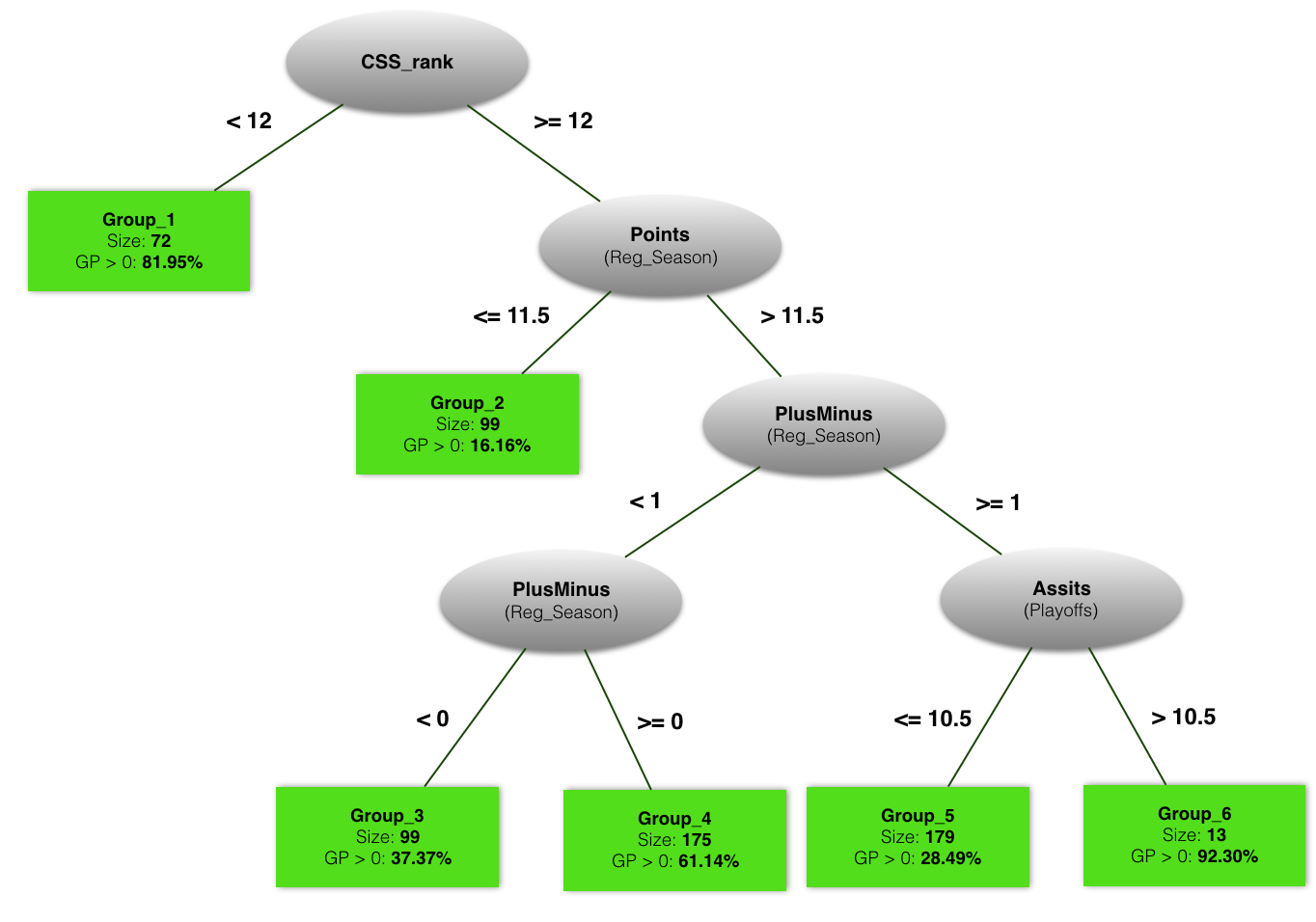


Figure 1 Logistic Regression Model Tree for the 2004,2005,2006 cohort. The Tree was built using the LogiBoost algorithm implemented in the LMT package of the Weka program [5, 6].

**5. Results: Predictive Modelling**

Following [1], we evaluated the predictive accuracy of the LMT model using the Spearman correlation between two player rankings: i) the performance ranking based on the actual number of NHL games that a player played, and ii) the ranking of players based on the probability *pi* of playing at least one game. We also compared it with iii) the ranking of players based on the order in which they were drafted. The draft order can be viewed as the ranking that reflects the judgement of NHL teams. We provide the formula for the Spearman correlation in the appendix. Table 3 shows the Spearman correlation for different rankings.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Training Data NHL Draft Years | Out of Sample  Draft Years | Draft Order Spearman Rank Correlation | Tree Model classification accuracy | Tree Model Spearman Rank Correlation |
| 1998, 1999, 2000 | 2001 | 0.43 | 82.27% | **0.83** |
| 1998, 1999, 2000 | 2002 | 0.30 | 85.79% | **0.85** |
| 2004, 2005, 2006 | 2007 | 0.46 | 81.23 % | **0.84** |
| 2004, 2005, 2006 | 2008 | 0.51 | 63.56 % | **0.71** |

Table 3 Predictive Performance (our model, over all draft ranking) using Spearman Rank Correlation. Bold indicates the best values.

*Other Approaches.* We also tried designs based on a linear regression model tree, using the M5P algorithm implemented in the Weka program. The result is a decision stump that splits on CSS rank only, which had substantially worse predictive performance than the logistic model tree. For the generalized additive model (*gam*), the reported correlations were 2001: 0.53, 2002: 0.54, 2007: 0.69, 2008: 0.71 [1]. Our correlation is not directly comparable to the *gam* model because of differences in data preparation: the *gam* model was applied only to drafted players who played at least one NHL game, and the CSS rank was replaced by the Cescin conversion factors: for North American players, multiply CSS rank by 1.35, and for European players, by 6.27 [7]. The Cescin conversion factors represent an interaction between the player’s country and the player’s CSS rank. A model tree offers another approach to representing such interactions: by splitting on the player location node, the tree can build a different model for each location. Whether the data warrant building different models for different locations is a data-driven decision made by the tree building algorithm. The same point applies to other sources of variability, for example the draft year or the junior league. In the next section we examine the interaction effects captured by the model tree in the different models learned in each leaf.

**6. Results: Learned Groups and Logistic Regression Models**

We examine the learned group regression models, first in terms of the dependent success variable, then in terms of the player features.

**6.1. Groups and the Dependent Variable**

Figure 2 shows boxplots for the distribution of our dependent variable *gi*. The strongest groups are, in order, 1, 6, and 4. The other groups show weaker performance on the whole, although in each group some players reach high numbers of games. Most players in Group 2&3&4&5 have GP equals to zero while Group 1&6 represent the strongest cohort in our prediction, where over 80% players played at least 1 game in NHL. The tree identifies that among the players who do not have a very high CSS rank (worse than 12), the combination of regular season Points > 11.5, PlusMinus > 0, and play-off Assists > 10 is a strong indicator of playing a substantive number of NHL games (median *gi* = 128).

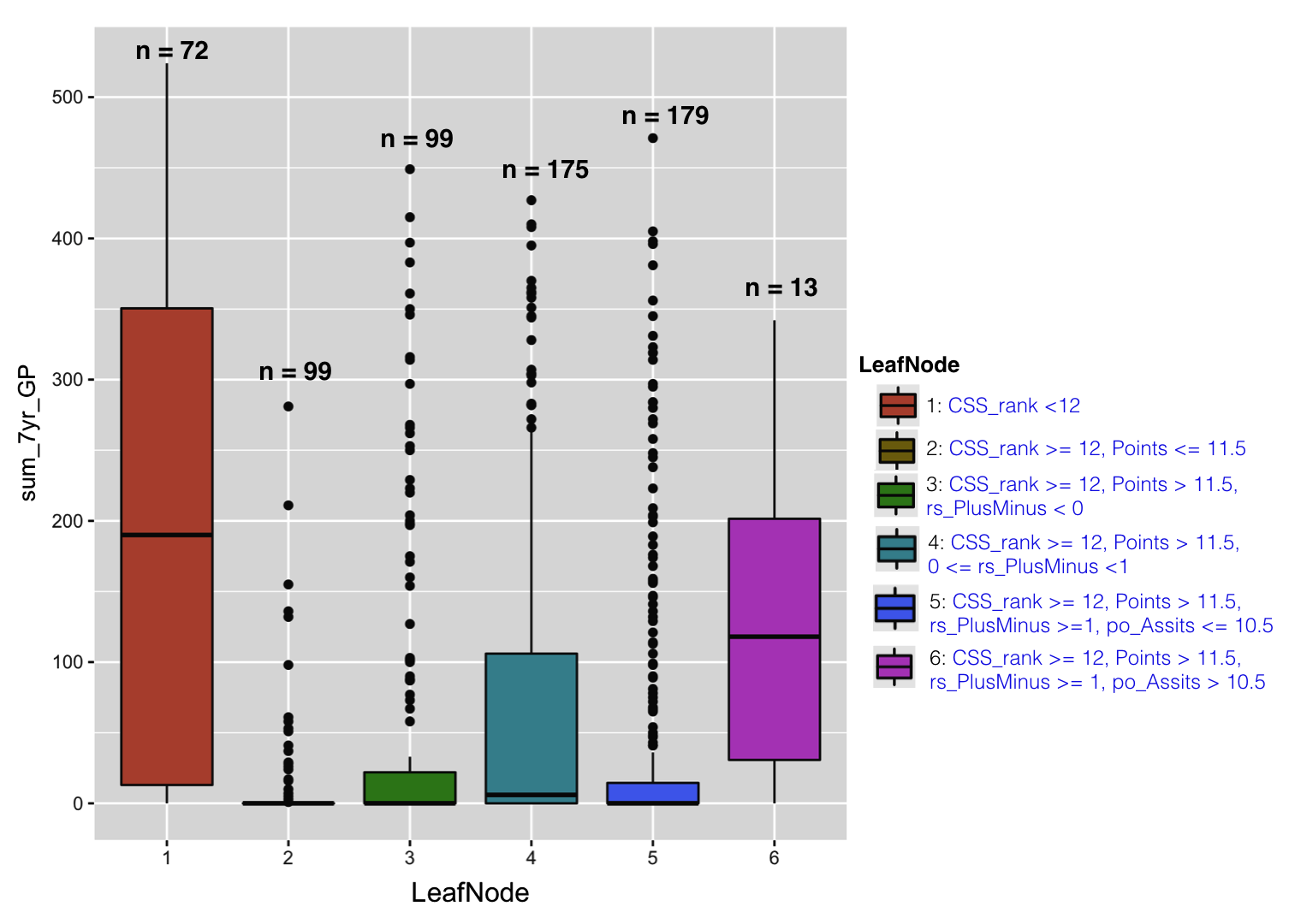


Figure 2 Boxplots for the dependent variable gi, the total number of NHL games played after 7 years under an NHL contract. Each boxplot shows the distribution for one of the groups learned by the logistic regression model tree. The group size is denoted n.

**6.2. Groups and the Independent Variables**

Table 4 shows the average statistics by group and for all players. The CSS rank for Group 1 is by far the highest. The data validate the high ranking in that 82% players in this group went on to play an NHL game. Group 6 in fact attains an even higher proportion of 92%. The average statistics of this group are even more impressive than those of group 1 (e.g. 67 regular season points in group 6 vs. 47 for group 1). But the average CSS rank is the lowest of all groups. So this group may represent a small group of players (*n* = 13) unappreciated by the scouts but identified by the tree. Other than Group 6, the group with the lowest CSS rank on average is Group 2. The data validate the low ranking in that only 16% of players in this group went on to play an NHL game. The group averages are also low (e.g. 6 regular season points is much lower than other groups).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group | Mean Points | | | | | | | | | | | |
| rs\_P | rs\_A | CSS\_  rank | po\_A | Weight | Height | rs\_  Plus  Minus | rs\_GP | rs\_  PIM | Draft  Age | po\_  GP | po\_P |
| 1 | 47 | 27 | 7 | 4 | 204 | 74 | 6 | 55 | 63 | 18 | 8 | 7 |
| 2 | 6 | 4 | 94 | 1 | 206 | 74 | -2 | 38 | 56 | 18 | 8 | 1 |
| 3 | 40 | 23 | 76 | 2 | 201 | 73 | -9 | 63 | 78 | 18 | 7 | 4 |
| 4 | 44 | 25 | 86 | 4 | 198 | 73 | 0 | 47 | 59 | 19 | 10 | 8 |
| 5 | 43 | 26 | 71 | 3 | 199 | 73 | 12 | 62 | 73 | 19 | 9 | 5 |
| 6 | 67 | 44 | 107 | 14 | 193 | 71 | 23 | 65 | 83 | 19 | 19 | 19 |
| *all* | *39* | *23* | *101* | *2* | *201* | *73* | *3* | *55* | *67* | *18* | *6* | *4* |

Table 4 Average statistics for each group and all players.

**7. Group Models and Variable Interactions**

Table 5 illustrates logistic regression weights by group. Bold numbers show the groups for which an attribute is most relevant. The table exhibits many interesting interactions among the independent variables; we have space to discuss only a few. Notice that *if the tree splits on an attribute, the attribute is assigned a high-magnitude regression weight by the logistic regression model for the relevant group*. Therefore our discussion focuses on the tree attributes.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Met-  rics  Group | CSS\_  rank | Draft  Age | Height | Weight | Country  \_group | Position | Games  \_played | Goals | Assists | Points | Penalty\_  in\_  Minutes | Plus  Minus |
| 1 | **-17.9** | -3.91 | -2.69 | 2.35 | E: -0.77  C: 1.23  U:-0.49 | D: -0.54  L: 2.09  R: -0.68 | rs: -2.43  po: 4.15 | rs: -0.03  po: **-9.8** | rs: 1.97  po: 8.89 | rs: 1.73  po: 0.3 | rs: **7.98**  po: **-7.6** | rs: -0.45 |
| 2 | -1.12 | -1.1 | -4.8 | **6.7** | E: -0.13  C: 0.28  U:-0.22 | D: -1.1  L: -0.45  R: 1.89 | rs: 5.9  po: -14.1 | rs: -2.17  po: -2.9 | rs: **11.8**  po: **21.6** | rs: **14.2**  po:**11.1** | rs: -2.72  po: 5.2 | rs: 1.57 |
| 3 | -2.6 | **6.95** | -7.4 | **6.7** | E: -2.4  C: 1.04  U: 2.34 | D: 0.39  L: 0.68  R: -0.4 | rs: 3.21  po: -1.05 | rs:-0.52  po: -0.6 | rs: 1.36  po: -0.39 | rs: 0.54  po:-2.6 | rs: -1.88  po: 2.7 | rs: 13.16 |
| 4 | -2.4 | 5.2 | -4.2 | -0.52 | E: 1.08  C: -0.40  U: -0.6 | D: -0.03  L: -0.24  R: 0.14 | rs: 3.58  po: -4.5 | rs: -2.16  po: 1.58 | rs: -0.12  po: 1.71 | rs: -1.4  po: 1.6 | rs: -2.72  po: 3.45 | rs: 0 |
| 5 | -0.65 | -3.89 | **0.01** | 4.68 | E: -1.26  C: 0.74  U: 0.47 | D: 0.91  L: -0.64  R: 0.05 | rs: 2.24  po: -0.25 | rs: 3.59  po: -1.7 | rs: -0.23  po: 0.33 | rs: 2.19  po:-0.8 | rs: -4.05  po: 6.86 | rs: -0.73 |
| 6 | -8.89 | 6.64 | -14.91 | 0.34 | E: -28.1  **C: 5.9**  U: 7.2 | D: 3.32  L: 0.74  R: -28.12 | **rs: 16.7**  po: 2.74 | rs: **21.6**  po: -9.7 | rs: -0.34  po: -0.43 | rs: -0.5  po:-0.4 | rs: 1.3  po: -1.6 | rs: **21.9** |

Table 5 Group 200(4 + 5 + 6 + 7 + 8) Weights Illustration. E = Europe, C = Canada, U = USA, rs = Regular Season, po = Playoff. Largest-magnitude weights are in bold. Underlined weights are discussed in the text.

At the tree root, *CSS rank* receives a large negative weight of -17.9 for identifying the most successful players in Group 1, where all CSS ranks are better than 12. Figure 3a shows that the proportion of above-zero to zero-game players decreases quickly in Group 1 with worse CSS rank. However, the decrease is not monotonic. Figure 3b is a scatterplot of the original data for Group 1. We see a strong linear correlation ( = -0.39), and also a large variance within each rank. The proportion aggregates the individual data points at a given rank, thereby eliminating the variance. This makes the proportion a smoother dependent variable than the individual counts for a regression model.

Group 5 has the smallest logistic regression coefficient of -0.65. Group 5 consists of players whose CSS ranks are worse than 12, regular season points above 12, and plus-minus above 1. Figure 4a plots CSS rank vs. above-zero proportion for Group 5. As the proportion plot shows, the low weight is due to the fact that the proportion trends downward only at ranks worse than 200. The scatterplot in Figure 4b shows a similarly weak linear correlation of -0.12.

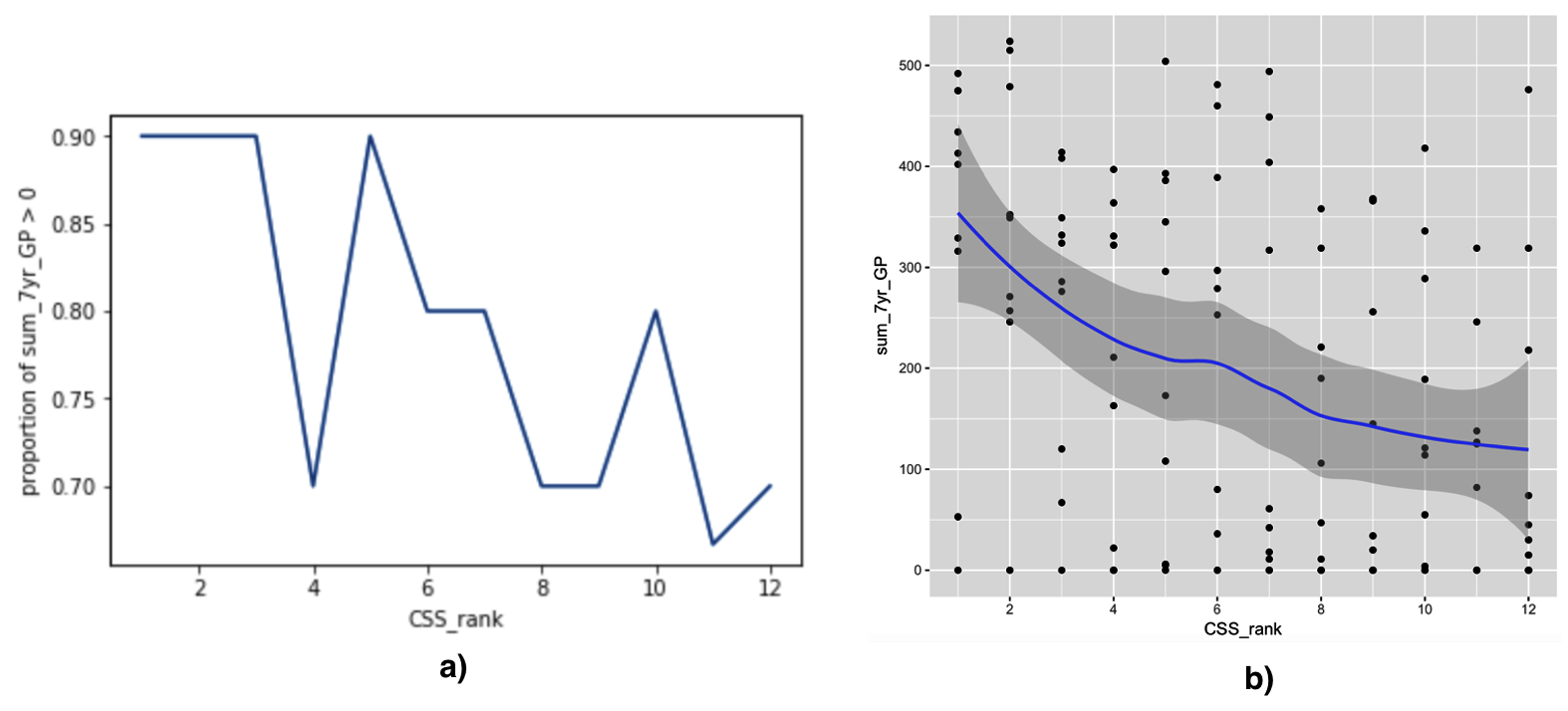


Figure 3 Proportion and scatter plots for CSS\_rank vs. sum\_7yr\_GP in Group 1.

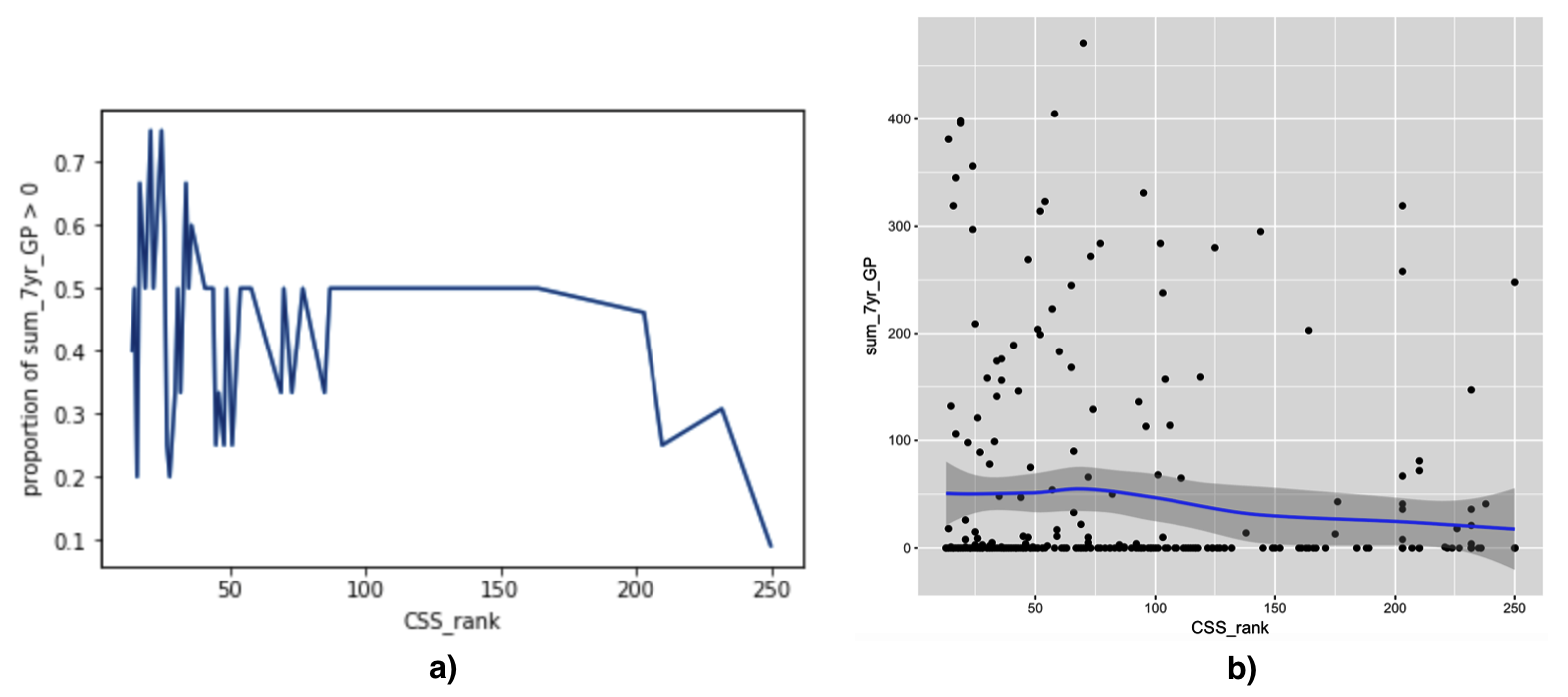


Figure 4 Proportion and scatter plots for CSS\_rank vs.sum\_7yr\_GP in Group 5.

*Regular season points* are the most important predictor for Group 2, which comprises players with CSS rank worse than 12, and regular season points below 12. In the proportion plot Figure 5, we see a strong relationship between points and the chance of playing more than 0 games (logistic regression weight 14.2). In contrast in Group 4 (overall weight -1.4), there is essentially no relationship up to 65 points; for players with points between 65 and 85 in fact the chance of playing more than zero games slightly decreases with increasing points.

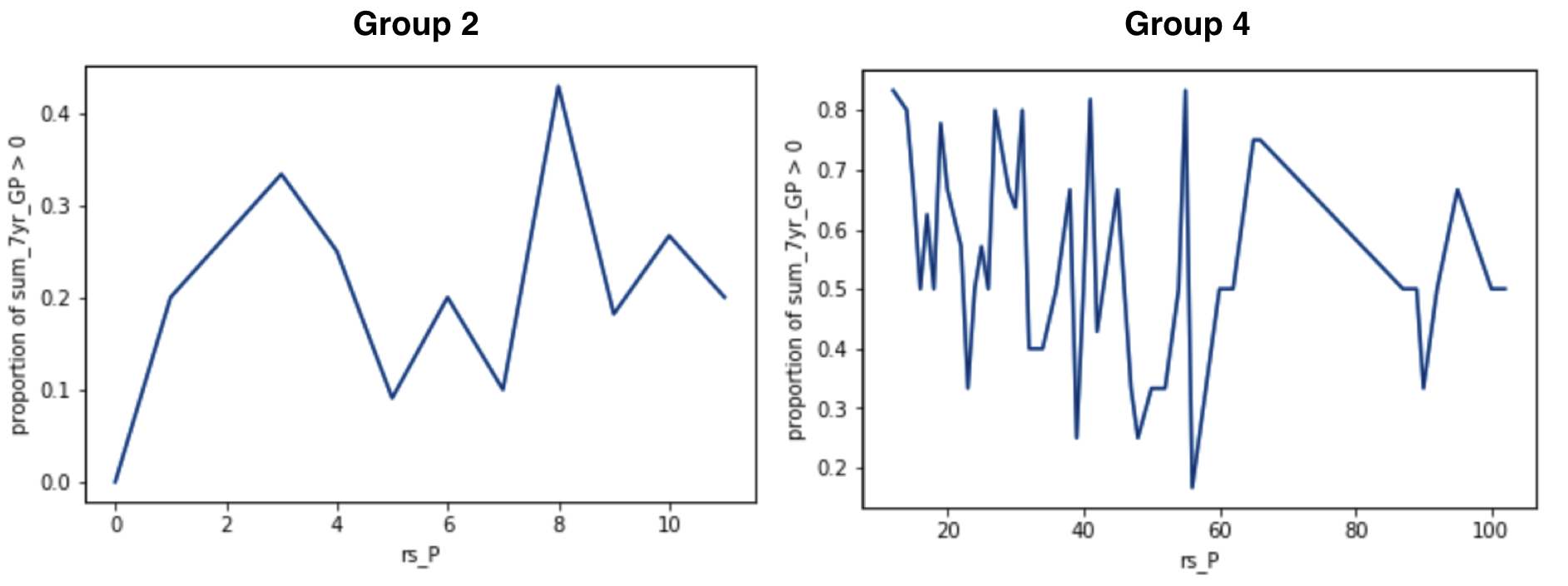


Figure 5 Proportion\_of\_Sum\_7yr\_GP\_greater\_than\_0 vs. rs\_P in Group 2&4.

In Group 3, players are ranked at level 12 or worse, have collected at least 12 regular season points, and show a negative plus-minus score. The most important feature for Group 3 is the *regular season plus-minus score* (logistic regression weight 13.16), which is negative for all players in this group. In this group, the chances of playing an NHL game increase with plus-minus, but not monotonically, as Figure 6 shows.

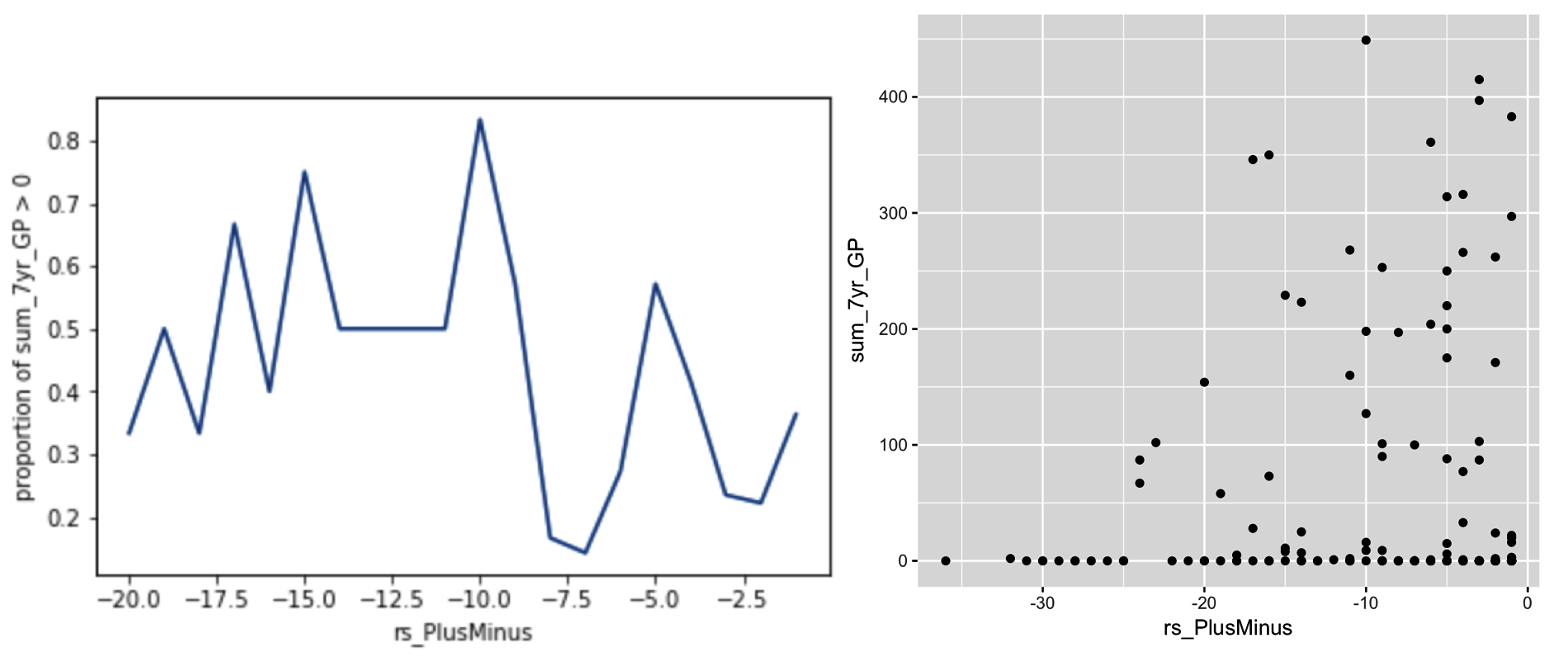


Figure 6 Proportion and scatter plots for rs\_PlusMinus vs.sum\_7yr\_GP in group 3.

For *regular season goals*, Group 5 assigns a high logistic regression weight of 3.59. Group 5 comprises players at CSS rank worse than 12, regular season points 12 or higher, and positive plus-minus greater than 1. About 64.8% in this group are offensive players (see Figure 7). The positive weight therefore indicates that successful forwards score many goals, as we would expect.

Group 2 assigns a surprising negative weight of -2.17 to the number of goals scored. Group 2 contains mainly defensemen (61.6%; see Figure 7). The typical strong defenseman scores 0 or 1 goals in this group. Players with more goals tend to be forwards, who are weaker in this group. In sum, the tree assigns weights to goals that are appropriate for different positions, using statistics that correlate with position (e.g. plus-minus), rather than the position directly.

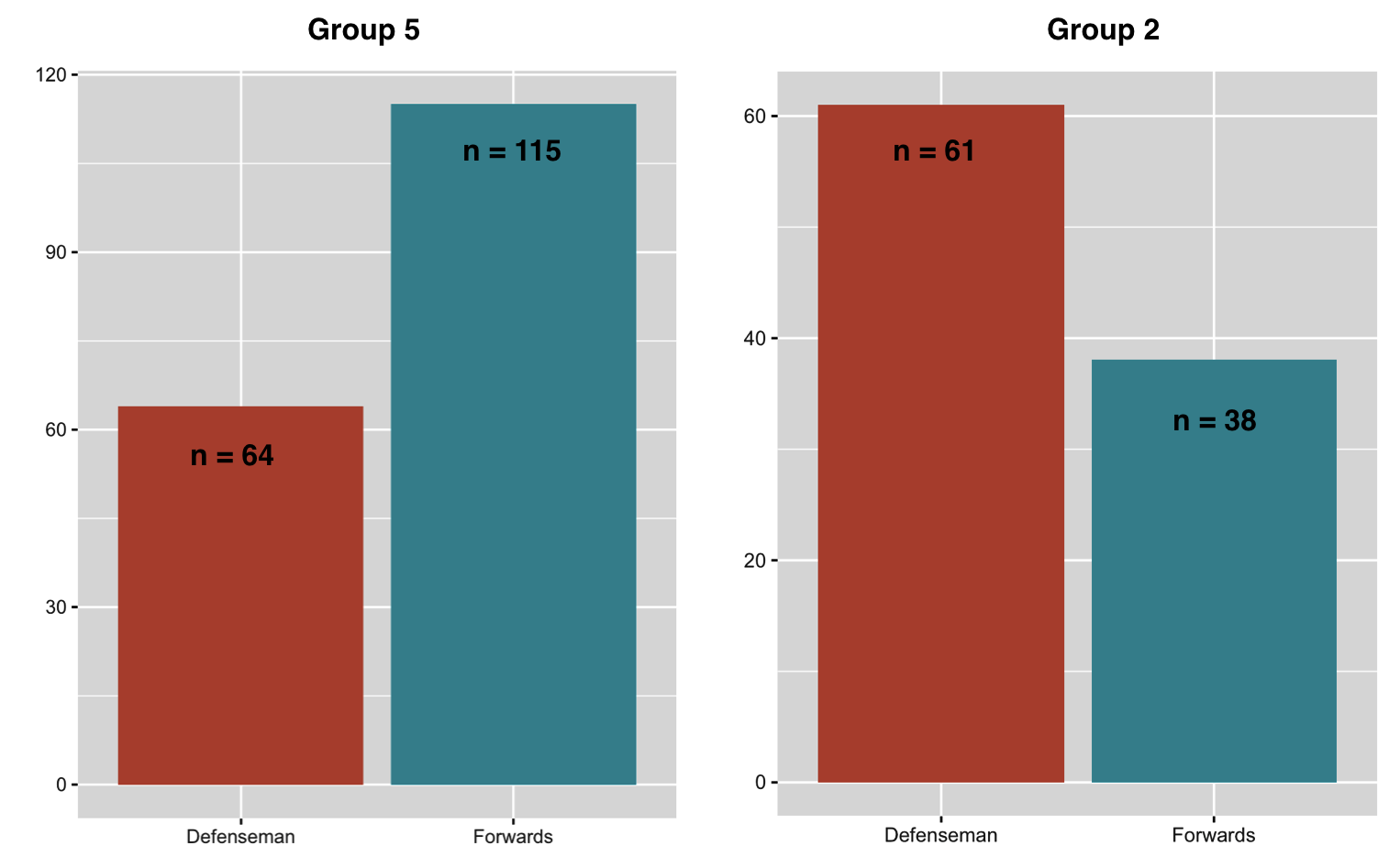


Figure 7 Distribution of Defenseman vs. Forwards in Group 5&2. The size is denoted as n.

**8. Identifying Exceptional Players**

Teams make drafting decisions not based on player statistics alone, but drawing on all relevant source of information, and with extensive input from scouts and other experts. As Cameron Lawrence from the Florida Panthers put it, “the numbers are often just the start of the discussion” [11]. In this section we discuss how the model tree can be applied to support the discussion of individual players by highlighting their special strengths. The idea is that the learned weights can be used to identify which features of a highly-ranked player differentiate him the most from others in his group.

**Explaining the Rankings: identify weak points and strong points**

Our method is as follows. For each group, we find the average feature vector of the players in the group, which we denote by *xg1, xg2,…, xgm* (see Table 4).We denote the features of player *i* as *xi1, xi2,…, xim.* Then given a weight vector *(w1,..,wm)* for the logistic regression model of group *g*, the log-odds difference between player *i* and a random player in the group is given by

We can interpret this sum as a measure of how high the model ranks player *i* compared to other players in his group. This suggests defining as the player’s *strongest features* the *xij* that maximize *wj(xij-xgj)*, and as his *weakest features* those that minimize *wj(xij-xgj).* This approach highlights features that are i) relevant to predicting future success, as measured by the magnitude of *wj*, and ii) different from the average value in the player’s group of comparables, as measured by the magnitude of *xij-xgj*.

**Case Studies.** Table 6 shows, for each group, the three strongest points for the most highly ranked players in the group. We see that the ranking for individual players is based on different features, even within the same group. The table also illustrates how the model allows us to identify a group of comparables for a given player. We discuss a few selected players and their strong points. The most interesting cases are often those where are ranking differs from the scouts’ CSS rank. We therefore discuss the groups with lower rank first.

Among the players who were not ranked by CSS at all, our model ranks *Kyle Cumiskey* at the top. Cumiskey was drafted in place 222, played 132 NHL games in his first 7 years, represented Canada in the World Championship, and won a Stanley Cup in 2015 with the Blackhawks. Given this record, arguably our model was correct to identify him as a strong NLH prospect. His strongest points were being Canadian, and the number of games played (e.g., 27 playoff games vs. 19 group average).

In the lowest CSS-rank group 6 (average 107), our top-ranked player *Brad Marchand* received CSS rank 80, even below his Boston Bruin teammate’s *Lucic*’s. Given his Stanley Cup win and success representing Canada, arguably our model was correct to identify him as a strong NLH prospect. The model highlights his superior play-off performance, both in terms of games played and points scored. Group 2 (CSS average 94) is a much weaker group*. Matt Pelech* is ranked at the top by our model because of his unusual weight, which in this group is unusually predictive of NHL participation. In group 4 (CSS average 86), *Sami Lepisto* was top-ranked, in part because he did not suffer many penalties although he played a high number of games. In group 3 (CSS average 76), *Brandon McMillan* is ranked relatively high by our model compared to the CSS. This is because in this group, left-wingers and shorter players are more likely to play in the NHL. In our ranking, *Milan Lucic* tops Group 5 (CSS average 71). At 58, his CSS rank is above average in this group, but much below the highest CSS rank player (*Legein* at 13). The main factors for the tree model are his high weight and number of play-off games played. Given his future success (Stanley Cup, NHL Young Stars Game), arguably our model correctly identified him as a star in an otherwise weaker group. The top players in Group 1 like *Sidney Crosby* and *Patrick Kane* are obvious stars, who have outstanding statistics even relative to other players in this strong group.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Top Players | Strongest Points (= group mean) | | |
| 1 | Sidney Crosby | rs\_P | rs\_A | CSS\_rank |
| 188 () | 110 () | 1 () |
| Patrick Kane | rs\_P | rs\_A | CSS\_rank |
| 154 () | 87 () | 2 () |
| Sam Gagner | rs\_P | po\_A | rs\_A |
| 118 () | 22 () | 83 () |
| 2 | Matt Pelech | Weight | CSS\_rank | rs\_A |
| 230 () | 41 () | 4 ( |
| Adam Pineault | CSS\_rank | rs\_P | Height |
| 25 () | 8 () | 73 () |
| Roman Wick | rs\_P | CSS\_rank | rs\_PlusMinus |
| 10 () | 36 () | 0 () |
| 3 | A.J.Jenks | CSS\_rank | Weight | Country |
| 20 () | 205 () | USA |
| Bill Sweatt | CSS\_rank | Position | rs\_PlusMinus |
| 27 () | L | -1 () |
| Brandon McMillan | CSS\_rank | Position | Height |
| 44 () | L | 71 () |
| 4 | Sami Lepisto | CSS\_rank | rs\_GP | rs\_PIM |
| 25 () | 61 () | 30 () |
| Linus Omark | CSS\_rank | Height | DraftAge |
| 55 () | 70 () | 20 () |
| Oscar Moller | CSS\_rank | Height | rs\_GP |
| 20 () | 70 () | 68 () |
| 5 | Milan Lucic | Weight | po\_GP | CSS\_rank |
| 236 () | 23 () | 58 () |
| Michael Del Zotto | Position | Country | po\_GP |
| D | CAN | 15 () |
| Steven Delisle | Weight | Country | po\_GP |
| 234 () | CAN | 19 () |
| 6 | Brad Marchand | Country | po\_GP | po\_P |
| CAN | 25 () | 23 () |
| Mathieu Carle | Country | CSS\_rank | rs\_GP |
| CAN | 53 () | 67 () |
| Kyle Cumiskey | Country | po\_GP | rs\_GP |
| CAN | 27 () | 72 () |

Table 6 Strongest Statistics for the top players in each group. Underlined players are discussed in the text.

**Conclusion and Future Work**

[best of both worlds, easy to obtain data]

Tree models are flexible and can be applied to other prediction problems to discover groups of comparable players as well as predictive models. For example, we can predict future NHL success from past NHL success, similar to Wilson [3] who used machine learning models to predict whether a player will play more than 160 games in the NHL after 7 years. Another direction is to apply the model to other sports, for example drafting for the National Basketball Association.

**References**

[1] Schuckers, M. & Statistical Sports Consulting, LLC. “Draft by Numbers: Using Data and Analytics to Improve National Hockey League (NHL) Player Selection.” MIT Sloan Sports Analytics Conference, 2016.

[2] Weissbock, J. *Draft Analytics: Unveiling The Prospect Cohort Success Model.* <https://canucksarmy.com/2015/05/26/draft-analytics-unveiling-the-prospect-cohort-success-model/>

[3] Wilson, D. “Mining NHL Draft Data and A New Value Pick Chart.” A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of Master of Science, 2016.

[4] <https://github.com/liuyejia/Model_Trees_Full_Dataset>

[5] Frank, E.; Hall, M. & Witten, I. *The WEKA Workbench.* Online Appendix for *Data Mining: Practical Machine Learning Tools and Techniques*. Fourth Edition, 2016.

[6] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I. “The WEKA Data Mining Software: An Update. SIGKDD Explorations.” Volume 11, Issue 1.

[7] Fyffe, I. *Evaluating Central Scouting*. Retrieved from <http://www.hockeyprospectus.com/puck/article.php?articleid=777>. (2011, January 19).

[8] Friedman, J.; Hastie, T & Tibshirani, R. “Additive Logistic Regression: A Statistical View of Boosting.” The Annals of Statistics, Vol.28, No.2, 337-407, 2000.

[9] Schuckers, M. & Argeris, S. “You Can Beat the Market: Estimating the Return on Investment for National Hockey League (NHL) Team Scouting using a Draft Value Pick Chart for the NHL.” <https://arxiv.org/abs/1411.5754> , 2014.

[10] Fieller, E. C.; Hartley, H. O. & Pearson, E. S. "Tests for rank correlation coefficients. I". *Biometrika*. **44**: 470-481. doi :[10.1093/biomet/44.3-4.470](https://doi.org/10.1093%2Fbiomet%2F44.3-4.470) , 1957.

[11] Joyce, E. & Lawrence, C. *Blending Old and New: How the Florida Panthers Try to Predict Future Performance at the NHL Entry Draft.* MIT Sloan Sports Analytics Conference, 2017.

[12] Cervone, D.; D’Amour, A.; Bornn, L. & Goldsberry, K. “POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data.” MIT Sloan Sports Analytics Conference, 2014.

[13] Thomas, A.; Ventura, S.; Jensen, S. & Ma, S. “Competing Process Hazard Function Models for

Player Ratings in Ice Hockey.” The Annals of Applied Statistics 7(3), 1497-1524, 2013.

[14] Schulte, O.; Zhao, Z. & SPORTLOGiQ. “Apples-to-Apples: Clustering and Ranking NHL Players

Using Location Information and Scoring Impact.” MIT Sloan Sports Analytics Conference, 2017.

[15] Macdonald B. “An improved Adjusted Plus-Minus Statistic for NHL Players”. MIT Sloan

Sports Analytics Conference, 2011.

[16] Gramacy, R.; Jesen, S. & Taddy, M. “Estimating Player Contribution in Hockey with

Regularized Logistic Regression.” J Quant Anal Sports 9:97-111, 2013.

[17] Albert, J.; Glickman, M. E.; Swartz, T. B. & Koning, R. H. *Handbook of Statistical Methods and Analyses in Sports*, CRC Press, 2013.

[18] Kaplan, E. H.; Mongeon, K. & Ryan, J. T. “A Markov model for hockey: Manpower differential and win probability added.” *INFOR: Information Systems and Operational Research* **52**(2), 39—50, 2014.

[19] Tingling, P.; Masri, K. & Martell M. “Does Order Matter? An Empirical Analysis of NHL Draft Decisions.” Sport, Business and Management: an International Journal. 1(2): 155-171, 2011.

[20] Cherkauer, K. J. “Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks.” In Chan, P. (Ed.), Working Notes of AAAI Workshop on Integrating Multiple Learned Model, pp. 15-21, 1996.

[21] Freund, Y. & Schapire, R. E. “Experiments with a new boosting algorithm.” In Proc. 13th International Conference on Machine Learning, pp. 148-146. Morgan Kaufmann.

[22] Hansen, L. & Salamon, P. “Neural Network Ensembles.” IEEE Trans. Pattern Analysis and Machine Intell., 12, 993-1001, 1990.

[23] Fraley, C. & Raftery, A.E. “Model-Based Clustering, Discriminant Analysis, and Density Estimation”. Journal of the American Statistical Association, 97, 611-631, 2002.

[24] Baudry, J.-P.; Raftery, A.E.; Celeux, G.; Lo, K. &Gottardo, R. “Combining Mixture Components for Clustering.” Journal of Computational and Graphical Statistics 19:332-353, 2010.

[25] Ryder, A. *Poisson Toolbox, a review of the application of the Poisson Probability Distribution in hockey*. Hockey Analytics, 2004.

[26] Loh, W.-Y. *GUIDE User Manual*, 2017. <http://www.stat.wisc.edu/~loh/guide.html>

[27] *NBA Draft Data Source*. <https://www.basketball-reference.com/draft/>

**Appendix**

**Spearman Rank Correlation**

Spearman’s correlation measures the relevance and direction of monotonic association between two variables [10]. The standard formula for calculating is based on the squared rank differences:

1. , formula for no tied ranks. n = number of ranks, = different in paired ranks. This is the formula we applied in Table 3.

In the case of tied ranks, an alternative is to use the Pearson correlation between ranks, computed by the formula:

1. , where *xi* = rank of player *i* according to ranking *x*, ditto for *yi*.

Players who have played zero NHL games are tied when ranked by the number of NHL games; this is the only case of ties. Table 7 repeats the calculation of Table 3 using the Pearson correlation among ranks (2) rather than the squared rank differences (1). With this measure it is still the case that the model ranking correlates more highly with actual number of games played than the team draft order.

|  |  |  |  |
| --- | --- | --- | --- |
| Training Data NHL Draft Years | Out of Sample  Draft Years | Draft Order Pearson Rank Correlation | Pearson Rank Correlation |
| 1998, 1999, 2000 | 2001 | 0.43 | **0.69** |
| 1998, 1999, 2000 | 2002 | 0.45 | **0.72** |
| 2004, 2005, 2006 | 2007 | 0.48 | **0.60** |
| 2004, 2005, 2006 | 2008 | 0.51 | **0.58** |

Table 7 Spearman correlation as Pearson correlation of ranks

**LogitBoost Algorithm**

Ensemble methods provide a combination of classifiers to obtain a better predictive result than any of the standalone constituents [20, 21, 22]. While a model tree constructs a set of different classifiers, it also partitions the space of players, so the prediction for each player is based on exactly one model only, rather than a weighted majority vote. The LogitBoost algorithm [8] combines a model tree with weighted votes by building a separate model for each tree node (including nonleaf nodes). The prediction for a specific player is then the weighted vote of all models along the branch assigned to the player. This offers some of the advantages of a hierarchical shrinkage model in smoothing parameter values so that predictions from players from similar groups tend to use similar weights. In this paper, we used the tree structure learned from LogitBoost, with simple maximum likelihood estimates for the weights to make the weight for a leaf more interpretable by fitting the data in the leaf’s group more closely. The GUIDE system [26] is a well-developed software package that supports building ensembles of model trees. While such an ensemble tends to have even higher predictive accuracy, we have in this paper built only a single model tree to maintain interpretability.