Model Trees for Identifying Exceptional Players in the NHL and NBA Drafts

Yejia Liu(✉), Oliver Schulte, and Chao Li

Department of Computing Science, Simon Fraser University, Burnaby, Canada
{yejial,oschulte,chao_li_2}@sfu.ca

Abstract. Drafting players is crucial for a team's success. We describe a data-driven interpretable approach for assessing prospects in the National Hockey League and National Basketball Association. Previous approaches have built a predictive model based on player features, or derived performance predictions from comparable players. Our work 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 or learned thresholds of features. Each leaf node in the tree defines a group of players, with its own 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 model tree shows better predictive performance than the actual draft order from teams' decisions. It can also be used to highlight the strongest points of players.

Keywords: Player ranking · Logistic Model Trees · M5 regression trees · National Hockey League · National Basketball Association

1 Introduction

Player ranking is one of the most studied subjects in sports analytics [1]. In this paper we consider predicting success in both the National Hockey League (NHL) and National Basketball Association (NBA) from pre-draft data, with the goal of supporting draft decisions. The publicly available pre-draft 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 soccer draft data. 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 NHL/NBA draft data take a regression approach or a similarity-based approach [7,19]. Regression approaches build a predictive model that takes as input a set of player features, such as demographics (age, height, weight) and pre-draft performance metrics (goals scored, minutes played), and output a predicted *success metric* (e.g. number

U. Brefeld et al. (Eds.): MLSA 2018, LNAI 11330, pp. 93–105, 2019.
https://doi.org/10.1007/978-3-030-17274-9_8

of games played, or player efficiency rating). Cohort-based approaches divide players into groups of comparables and predict future success based on a player's cohort. For example, the PCS model [28] 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, have been developed to identify groups of comparables for each player. Our aim in this paper is to describe a new model for draft data that combines the advantages of both approaches, regression-based and similarity-based. Our method uses a model tree [4,13]. 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. Figure 1 shows an example model tree. A model tree offers several advantages.

- Compared to a single regression model, the tree defines *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.
- Compare to a similarity-based model, *tree construction learns groups of players from the data*, without requiring the analyst to specify a similarity metric. Because tree learning selects splits that increase predictive accuracy, the learned distinctions between the groups are guaranteed to be predictively relevant to a player's future career success. Also, the tree creates a model, not a single prediction, for each group, which allows it to differentiate players from the same group.

In the NHL draft, we approach prospect ranking not by directly predicting the future number of games played, but by *predicting whether a prospect will play any number of games at all in the NHL*, due to an excess-zeros problem [11]. For the NBA draft, we use a linear regression model tree that directly predicts the continuous success variable.

Following the work of Schuckers et al. and Greene [7,19], we evaluate the model trees ranking results by comparing to a ranking based on the players' actual future success, measured as the number of career games they played after 7 years in the NHL, or player efficiency rating (PER) in the NBA. The rank correlation of our logistic regression ranking for the NHL draft is competitive with that achieved by the generalized additive model of [19], which is currently the state-of-the-art for NHL draft data. As for the NBA draft, our linear regression ranking performs better than the actual draft pick in terms of correlations with players' future success.

We show in case studies 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 that scouts and teams can take into account in their evaluation.

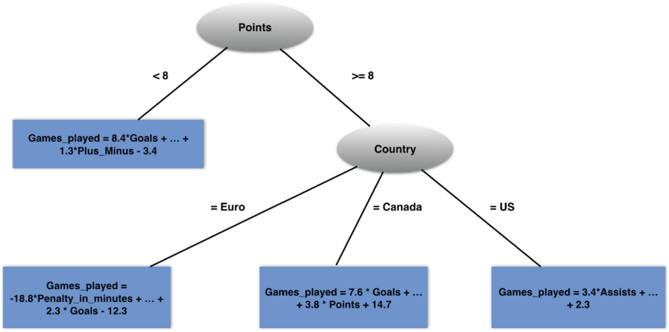
Points
< 8
>= 8
Games_played = 8.4*Goals + ... + 1.3*Plus_Minus - 3.4
Country
= Euro
= Canada
= US
Games_played = -18.8*Penalty_in_minutes + ... + 2.3 * Goals - 12.3
Games_played = 7.6 * Goals + ... + 3.8 * Points + 14.7
Games_played = 3.4*Assists + ... + 2.3

Fig. 1. A simple model tree example.

2 Related Work

Different approaches to player ranking are appropriate for different data types. For example, with dynamic play-by-play data, Markov models have been used to rank players [2,20,24]. For data that record the presence of players when a goal is scored, regression models have also been applied to extend the classic wins-contribution metrics [15,22]. In this paper, we utilize player statistics that aggregate a season's performance into a single set of numbers. While these data are much less informative than play-by-play data, they are easier to obtain, interpret, and process.

Regression Approaches. To our knowledge, this is the first application of model trees to hockey draft prediction, and the first model for predicting whether a draftee plays any games at all. The closest predecessor to our work is due to Schuckers [19] and Greene [7], who use a single linear regression model to predict future success from junior league data.

Similarity-Based Approaches assume a similarity metric and group similar players to predict performance. A sophisticated example from baseball is the nearest neighbour analysis in the PECOTA system [23]. For ice hockey, the Prospect Cohort Success (PCS) model [28], 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 whether a player will score more than 200 games career total. Tree learning can easily be modified to make predictions for any game count threshold. For basketball, many clustering approaches focus on defining appropriate roles or positions for a player. In Lutz's work [14], NBA players are clustered to several types like Combo Guards, Floor Spacers and Elite Bigs. The sports analytics group of Yale University has also developed an NBA clustering system to cluster players through hierarchical clustering methodology with their season performance statistics as inputs [6]. Our model tree also identifies which positions differ with respect to the statistical relationships between draft data and future performance.

Archetypoid analysis is a sophisticated approach to clustering groups of comparable players and identifying exceptional ones. Unlike our approach, it does not build a predictive model [27].

3 Datasets

We describe our dataset sources and preprocessing steps.

Ice Hockey Data. Our data were obtained from public-domain on-line sources, including nhl.com, eliteprospects.com, and thedraftanalyst.com. We are also indebted to David Wilson for sharing his NHL performance dataset [29]. The full dataset is posted on the worldwide web[1]. We consider players drafted into the NHL between 1998 and 2008 (excluding goaies) and divided them into two cohorts (1998–2002 cohort and 2004–2008 cohort). Our data include demographic factors (e.g. age, weight, height), performance metrics for the year in which a player was drafted (e.g., goals scored, plus/minus), career statistics (e.g. number of games played, time on ice), 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. If a player played for more than one team in his draft year (e.g., a league team and a national team), we added up the counts from different teams. We also eliminated players drafted in 2003 since most of them have no CSS rank [19].

Basketball Data. Our basketball datasets were obtained from https://www.basketball-reference.com, a rich resource of NBA player data, containing both pre-draft and career information. We posted our data in the worldwide web[2]. We considered players drafted into NBA between 1985 and 2011. Our training data included statistics of players drafted from 1985 to 2005, while players drafted from 2006 to 2011 were considered as our testing data. We excluded players whose college performance statistics are not available. For the 15 drafted players whose career statistics are not available, we replaced their career PER by $min(x) - std(x)$, where $min(x)$ is the minimum career PER and $std(x)$ is the standard deviation of career PER for all players drafted in the same year. The motivation is that these players are very likely judged to be worse, by coaches and team experts, than players who played in the NBA. We leave other imputation methods for future work.

[1] https://github.com/liuyejia/Model_Trees_Full_Dataset.

[2] https://github.com/sfu-cl-lab/Yeti-Thesis-Project/tree/master/NBA_work.

4 Methodology

4.1 Success Metrics

In the NHL draft, we took as our dependent variable *the total number of games g_i 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 [29]. Other interesting success metrics like the Wins Above Replacement (WAR) or Total Hockey Rating (ThoR) usually require play-by-play data and are often computationally expensive [18,25]. For the NBA, we adopted *player efficiency rating (PER)* as our success metric, which encompasses a large set of inputs and takes nearly every aspect of a player's contribution into consideration. The calculation is as follows:

$$PER = (uPER \times \frac{lg\,Pace}{tm\,Pace}) \times \frac{15}{lg\,uPER}$$

where $uPER$ is the unadjusted PER, calculated using player performance variables, as well as team and league statistics (e.g. game pace). In the above formula, lg is indicates the league, tm the team [9]. PER is widely used, but other metrics have been proposed [21]. Model trees can be applied with any performance metric.

4.2 Model Trees

Model Trees are a flexible formalism that can be built for any regression model. We use a logistic regression model for the NHL and linear regression for the NBA.

Logistic Model Tree (NHL). Our logistic regression model tree *predicts whether a player will play any games at all in the NHL* ($g_i > 0$). The motivation is that many players in the draft never play any NHL games at all (up to 50% depending on the draft year) [26]. This poses an extreme *excess-zeros problem* for predicting directly the number of games played. In contrast, for the classification problem of predicting whether a player will play any NHL games, excess-zeros means that the dataset 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 $p_i = P(g_i > 0)$ 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. We built a model tree whose leaves contain a logistic regression model, where each player can be assigned to a unique leaf node to compute a probability p_i.

Figure 2 shows the logistic regression model tree learned for our second cohort (players drafted between 2004–2008). 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

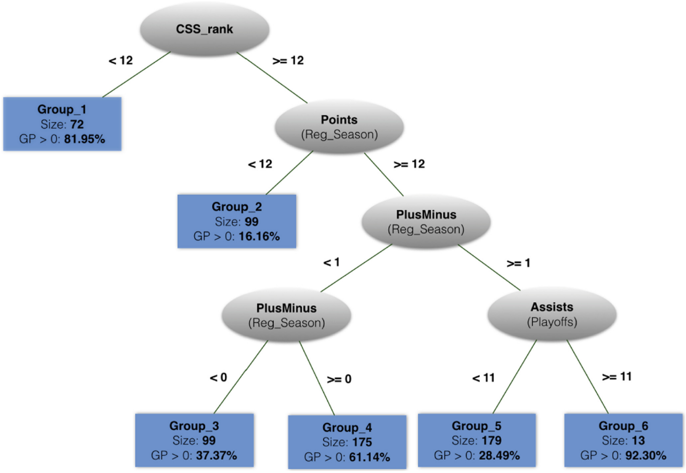

Fig. 2. Logistic Regression Model Trees (LMT) for the 2004, 2005, 2006 cohort in NHL. The tree was built using the LogitBoost algorithm implemented in the LMT package of the Weka Program [3,8]. Each leaf defines a group of players. For each group, the figure shows the proportion of players who played at least one game in the NHL. Each leaf contains a logistic regression model for its group (not shown), which produces different predictions for players from the same group but with different features. CSS_rank denotes rankings from the Central Scouting Service; lower rank numbers are better (e.g. rank 1 is the best).

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 *playoffs assists* as the next attribute. Players with a positive plus/minus score and more than 10 playoffs assists form a small but strong group that is 92% likely to play at least one NHL game.

Linear Regression Tree (NBA). Different from NHL, most drafted basketball players played at least one game in the NBA (over 80% in our datasets depending on the draft year). Since there is no excess-zeros issue, we used linear regression, which links predictors to the continuous dependent variable directly. Similar to the construction of logistic model tree in the NHL draft, we built a linear regression tree whose leaves contain a linear regression model. Each player can be assigned to its own leaf node to predict his future career PER.

Figure 3 shows our learned M5 regression model tree [16]. The root attribute is *position*, as the most important attribute due to its highest information gain.

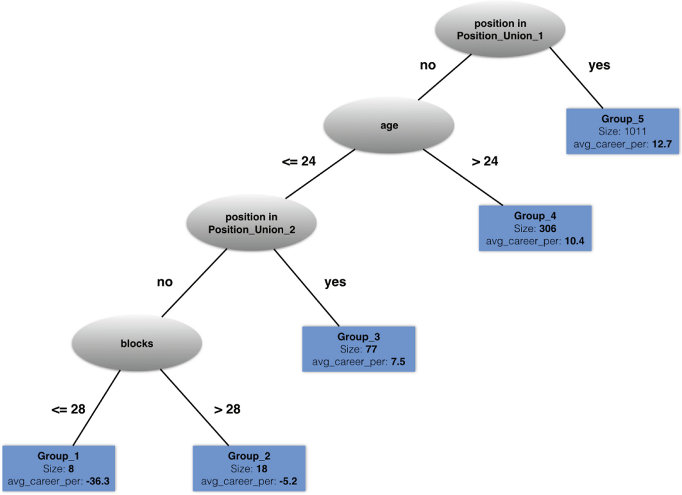

Fig. 3. M5 regression trees for the drafted players in 1985–2011 drafts. The tree was built with the M5P package of Weka [3,8]. Each leaf defines a group of players, and a linear regression model for its group (not shown). For each group, the figure shows the average career PER. The group model assigns weights to *all* player features, and produces different predictions for players from the same group with different features. The values of Position_Union_1 and Position_Union_2 are listed in [12].

Players who are from Position_Union_1 have an average PER about 13, forming a better group compared to other players. Position_Union_1 is a list of player positions (e.g. Center), automatically grouped by the M5P algorithm. For players who are not from Position_Union_1, the tree takes *age* as the next splitting attribute. Players who are older than 24 years old and are not from Position_Union_1, belong to a less promising group with PER around 10. Then, the tree chooses *position* as another splitting point again, reflecting its significance. For players who do not belong to Position_Union_1 but belong to Position_Union_2, with age smaller than or equal to 24, they form an average level group. The average PER value of players in this group is around 7. Position_Union_2 is another list of positions, which were automatically grouped by the M5P algorithm. Lastly, the tree chooses *blk (blocks)* as the splitting feature. The sizes of Group 1 and Group 2 are relatively small (8 and 18). These are special groups that require a customized model, according to tree. The tree finds conditions that separate the strong group 5 from the groups 1 and 2 that show substantially worse average performance.

For discrete variables like Position, the M5P algorithm has the capacity to evaluate splits based not only on specific values (e.g. Position = Center), but also on a disjunction of values (e.g. Position = Center or Forward). The disjunctions of positions represented by Position_Union_1 and Position_Union_2 offered the best trade-off between model complexity and data fit. Moreover, in its leaf model, the tree groups the positions further into subgroups (3 for Group 5 and 4 for

Group 4), which defines a shallow hierarchy of positions. Finding a set of position types is a much discussed question in basketball analytics [5,14,21] into position types. There has been considerable discussion of what player types are useful in basketball [6,14]. The model tree learns two new position types that are the union of previously introduced types.

In terms of the learned linear models, most players (about 72% are assigned to Group 5. Figure 4 shows the linear model defined for this group. For contrast, the Figure shows the linear model for the next-biggest group 4. Inspection shows that the weights differ substantially, which justifies the use of a model tree rather than a single model. Overall, we can view the linear model tree as defining one standard model for most players in Group 5, and a number of more specific models for special cases defined by the other groups.

Metrics Group	age	position	g	mp	ft	fta	trb	ast	blk	pts	ah
4	-34.35	1.92 0.05 0.07	12.89	all: 0 per: 5.08	all: 2.25 per: -0.14	-18.63	all: 0.21 per: 0	all: 0 per: 0.11	9.51	all: -11.24 per: 17.91	0.04
5	-2.39	0.36 0.83 0.53 1.34	-6.54	all: 4.27 per: -4.19	all: 10.57 per: -5.33	-10.69	all: 10.05 per: -4.95	all: 5.66 per: 0.04	2.41	all: 2.92 per: 4.01	1.03

Fig. 4. Example of learned weights. The table shows the weights of each feature in the linear equation for Group 4 and Group 5. *Per* represents per game statistics and *all* denotes overall statistics. g = games played, mp = minutes played, ft = free throws, fta = free throws attempt, trb = total rebounds, ast = assists, blk = blocks, pts = points, ah = amature honor. For possible values of Position, see the text.

5 Results

5.1 Modelling Results of the NHL Draft

Following [19], we consider three rankings as follows:

1. The performance ranking based on the actual number of NHL games that a player played.
2. The ranking of players based on the probability p_i of playing at least one game (Tree Model SRC).
3. The ranking of players based on the order in which they were drafted by team (Draft Order SRC).

Table 1 gives results for the out of sample prediction for players drafted in each of the four out of sample drafts using *games played* (GP) as the response variable. The draft order can be viewed as the ranking that reflects the judgment of NHL teams. Like [19], we evaluate the predictive accuracy of the Draft Order and the LMT model using the Spearman Rank Correlation (SRC) between (i)

Draft order and actual number of games, and (ii) p_i order and actual number of games, as shown in Table 1. We can see from Table 1 that the average rank correlation between the NHL draft order and the NHL Performance metric is about 0.43 while our tree model averages about 0.81 for both cohorts. This strongly suggests that our model outperforms the draft ordering by player selection.

Table 1. Predictive Performance (Draft Order, our Logistic Model Trees) measured by Spearman Rank Correlation (SRC) with actual number of games. Bold indicates the best values.

Training data NHL draft years	Out of sample draft years	Draft Order SRC	LMT classification accuracy	LMT SRC
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**

5.2 Modelling Results of the NBA Draft

We used both the Pearson Correlation and the Spearman Rank Correlation to compare the predictive power of our tree models to the actual draft order and a baseline method (Ordinary Linear Regression). As shown in Table 2, our model tree performs better than the actual draft order and ordinary linear regression.

Table 2. Comparison of predictive performance between draft order, linear regression and our tree models. Bold indicates the best values.

Method	Evaluation		
	Pearson Correlation	Spearman Rank Correlation	RMSE
Draft order	0.42	0.39	NaN
Ordinary linear regression	0.45	0.40	7.14
Our Model Tree	**0.55**	**0.43**	**6.16**

6 Case Studies: Exceptional Players and Their Strong Points

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” [10]. 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.

6.1 Explaining the Rankings: Identify 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 $(\overline{x_{g_1}}, \overline{x_{g_2}}, ..., \overline{x_{g_m}})$. We denote the features of player i as $(x_{i_1}, x_{i_2}, ..., x_{i_m})$. Then given a weight vector $(w_1, ..., w_m)$ 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

$$\sum_{j=1}^{m} w_j (x_{i_j} - \overline{x_{g_i}})$$

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 x_{i_j} that maximize $w_j(x_{i_j} - \overline{x_{g_i}})$. This approach highlights features that are (i) relevant to predicting future success, as measured by the magnitude of w_j, and (ii) different from the average value in the player' s group of comparables, as measured by the magnitude of $x_{i_j} - \overline{x_{g_i}}$. Table 3 summarizes the strongest players and their strongest statistics in the NHL and NBA draft.

Table 3. Strongest statistics for the top three players in the strongest group for the NHL and NBA draft [4].

Top players	Strongest points ($\overline{x}$ = group mean)		
NHL			
Sidney Crosby	Points *(regular season)* 188 ($\overline{x}$ = 47)	Assists *(regular season)* 110 ($\overline{x}$ = 27)	CSS_rank 1 ($\overline{x}$ = 7)
Patrick Kane	Points *(regular season)* 154 ($\overline{x}$ = 47)	Assists *(regular season)* 87 ($\overline{x}$ = 27)	CSS_rank 2 ($\overline{x}$ = 7)
Sam Gagner	Points *(regular season)* 118 ($\overline{x}$ = 47)	Assists *(playoffs)* 22 ($\overline{x}$ = 4)	Assists *(regular season)* 83 ($\overline{x}$ = 27)
NBA			
Larry Johnson	Free throws 162 ($\overline{x}$ = 122)	Total rebounds 380 ($\overline{x}$ = 214)	Assists 104 ($\overline{x}$ = 84)
Anfernee Hardaway	Total rebounds 273 ($\overline{x}$ = 214)	Assists 204 ($\overline{x}$ = 84)	Minutes played 1196 ($\overline{x}$ = 929)
Chris Webber	Total rebounds 362 ($\overline{x}$ = 84)	Minutes played 1143 ($\overline{x}$ = 929)	Assists 90 ($\overline{x}$ = 84)

6.2 Case Studies

NHL Draft. *Sidney Crosby* and *Patrick Kane* are obvious stars, who have outstanding statistics even relative to other players in this strong 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.

The most interesting cases are often those whose ranking differs from the scouts' CSS rank. 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. His *strongest points* were being Canadian, and the number of games played (e.g., 27 playoffs games vs. 19 group average). In the lowest CSS-rank group, Group 6, our top-ranked player *Brad Marchand* received CSS rank 80, even below his Boston Bruin teammates Lucic's. Given his Stanley Cup win and success representing Canada, arguably our model was correct to identify him as a strong NHL prospect. The model *highlights* his superior play-off performance, both in terms of games played and points scored.

NBA Draft. The most prestigious player *Chris Webber* identified by our model was a superstar in the NBA. He is a five-time NBA All-Star, a five-time All-NBA Team member, and NBA Rookie of the Year. His *strongest points* in his pre-draft year are trb (total rebounds), mp (minutes played) and ast (assists). There are also examples of players highlighted by our model, who in hindsight were undervalued compared to other players from the same cohort. For instance, *Matt Geiger* was ranked at 13*th* in his group by our model tree, who was picked at 42*th* in the draft, after Todd Day (8th) and Bryant Stith (13th). However, his career PER is 15.2, above these two players drafted before him. His pre-draft *strongest points* were identified as total rebounds, assists and minutes played. A more recent instance is *Dejuan Blair*, who had the 37*th* overall draft pick in 2009, taken after Jordan Hill (8th), Ricky Rubio (5th), but he obtained almost the same career PER as them.

7 Conclusion

We have proposed building regression model trees for ranking draftees in the NHL and NBA, or other sports, based on a list of player features and performance statistics. The model tree groups players according to the values of discrete features, or learned thresholds for continuous performance statistics. Each leaf node defines a group of players that is assigned its own regression model. Tree models combine the strength of both regression and cohort-based approaches, where player performance is predicted with reference to comparable players.

Key findings include the following: (1) The model tree ranking correlates well with the actual success ranking according to the actual number of games played, better than draft order. (2) The model tree can highlight the exceptionally strong points of draftees that make them stand out compared to the other players in their group.

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 [29] 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 Football League.

Future Work. A common issue in drafting is comparing players from different leagues. A model tree offers a promising approach to this issue as it can make data-driven decisions about whether players from different leagues should be assigned to different models. For example, preliminary experiments with the NHL data show that adding the junior league of a player as a feature leads, the tree to split on the player's country. This finding suggests that model tree learning can be applied to assess the skills of both North American and international NBA prospects in a single model. Identifying which skills of international players make them compatible with the NBA has become increasingly important over the past decades [17]. Building a model tree to make up-to-date predictions for the current draft would be the most relevant application for teams.

References

1. Albert, J., Glickman, M.E., Swartz, T.B., Koning, R.H.: Handbook of Statistical Methods and Analyses in Sports. CRC Press, Boca Raton (2017)
2. Cervone, D., D'Amour, A., Bornn, L., Goldsberry, K.: POINTWISE: predicting points and valuing decisions in real time with NBA optical tracking data. In: MIT Sloan Sports Analytics Conference (2016)
3. Frank, E., Hall, M., Witten, I.: The WEKA workbench. Online appendix for data mining: practical machine learning tools and techniques (2016)
4. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Ann. Stat. **28**(2), 337–407 (2000)
5. Green, E., Menz, M., Benz, L., Zanuttini-Frank, G., Bogaty, M.: Clustering NBA players (2016). https://sports.sites.yale.edu/clustering-nba-players
6. Green, E., Menz, M., Benz, L., Zanuttini-Frank, G., Bogaty, M.: Clustering NBA players (2017). https://sports.sites.yale.edu/clustering-nba-players
7. Greene, A.C.: The success of NBA draft picks: can college career predict NBA winners. Master's thesis, St. Cloud State University (2015)
8. Hall, M., Frank, E., Homes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: an update. SIGKDD Explor. **11**, 10–18 (2009)
9. Hollinger, J.: Calculating PER (2013). https://www.basketball-reference.com/about/per.html
10. Joyce, E., Lawrence, C.: Blending old and new: how the Florida panthers try to predict future performance at the NHL entry draft (2017)
11. Lachenbruch, P.A.: Analysis of data with excess zeros. Stat. Methods Med. Res. **11** (2002)

12. Liu, Y., Schulte, O., Hao, X.: Drafting NBA players based on their college performance (2018). https://github.com/liuyejia/Model_Trees_Full_Dataset/blob/master/NBA_work/ReadMe_NBA.md
13. Loh, W.Y.: GUIDE user manual. University of Wisconsin-Madison (2017)
14. Lutz, D.: A cluster analysis of NBA players. In: MIT Sloan Sports Analytics Conference (2012)
15. Macdonald, B.: An improved adjusted plus-minus statistic for NHL players. In: MIT Sloan Sports Analytics Conference (2011)
16. Quinlan, J.R.: Learning with continuous classes, pp. 343–348. World Scientific (1992)
17. Salador, K.: Forecasting performance of international players in the NBA. In: MIT Sloan Sports Analytics Conference (2011)
18. Schuckers, M., Curro, J.: Total Hockey Rating (THoR): a comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. In: MIT Sloan Sports Analytics Conference (2013)
19. Schuckers, M.E., Statistical Sports Consulting, LLC: Draft by numbers: using data and analytics to improve National Hockey League (NHL) player selection. In: MIT Sloan Sports Analytics Conference (2016)
20. Schulte, O., Zhao, Z., SPORTLOGiQ: Apples-to-apples: clustering and ranking NHL players using location information and scoring impact. In: MIT Sloan Sports Analytics Conference (2017)
21. Shea, S., Baker, C.: Basketball Analytics: Objective and Efficient Strategies for Understanding How Teams Win. CRC Press, Boca Raton (2017)
22. Sill, J.: Improved NBA adjusted +/− using regularization and out-of-sample testing. In: MIT Sloan Sports Analytics Conference (2010)
23. Silver, N.: PECOTA 2004: A Look Back and a Look Ahead, pp. 5–10. Workman Publishers, New York (2004)
24. Thomas, A., Ventura, S., Jensen, S., Ma, S.: Competing process hazard function models for player ratings in ice hockey. Ann. Appl. Stat. **7**(3), 1497–1524 (2013)
25. Thomas, A.C., Ventura, S.: The highway to WAR: defining and calculating the components for wins above replacement (2015). https://aphockey.files.wordpress.com/2015/04/sam-war-1.pdf
26. Tingling, P., Masri, K., Martell, M.: Does order matter? An empirical analysis of NHL draft decisions. Sport Bus. Manag.: Int. J. **1**(2), 155–171 (2011)
27. Vinué, G., Epifanio, I.: Archetypoid analysis for sports analytics. Data Min. Knowl. Discov. **31**(6), 1643–1677 (2017)
28. Weissbock, J.: Draft analytics: unveiling the prospect cohort success model. Technical report (2015). https://canucksarmy.com/2015/05/26/draft-analytics-unveiling-the-prospect-cohort-success-model/
29. Wilson, D.R.: Mining NHL draft data and a new value pick chart. Master's thesis, University of Ottawa (2016)