Authors: Oliver Schulte, Yejia Liu, Chao Li, School of Computing Science, Simon Fraser University, Vancouver, Canada

Model Trees for Identifying Exceptional Players in the NHL Draft

Introduction

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.

Methods

Following [1], our dependent variable is the number g_i of games that player i plays within his first 7 years of NHL play. Model trees are flexible and can be employed with any regression model. In this paper we employ logistic regression, where the dependent binary variable is 1 if $g_i > 0$. We applied the standard LMT algorithm. To each player i, the learned model tree assigns a probability $p_i(g_i > 0)$ of playing more than 0 games. Players are ranked according to p_i. Players with $p_i < 0.5$ are assigned to the zero-count bottom rank, which offers a new approach to the zero-inflation problem: almost half the draftees never play an NHL game [1].

To assess predictive accuracy as in [1], players are ranked by i) the learned model ii) the actual number of games g_i, iii) the actual draft order. For our ranking i vs. ii, we have Spearman rank correlation $r = 0.81$. For ranking ii vs. iii, $r = 0.48$.
Table 2 illustrates the learned clusters and shows the top player in each cluster. We extract for each player their strongest feature that contributes the most to increasing their predictive score.

Results

Figure 1 shows the model tree learned on our data set. We crawled the data from on-line sources (nhl.com, eliteprospects.com, thedraftanalyst.com), and posted the set at https://github.com/liuyejia/Model_Trees_Full_Dataset. Table 1 illustrates the two strongest groups.

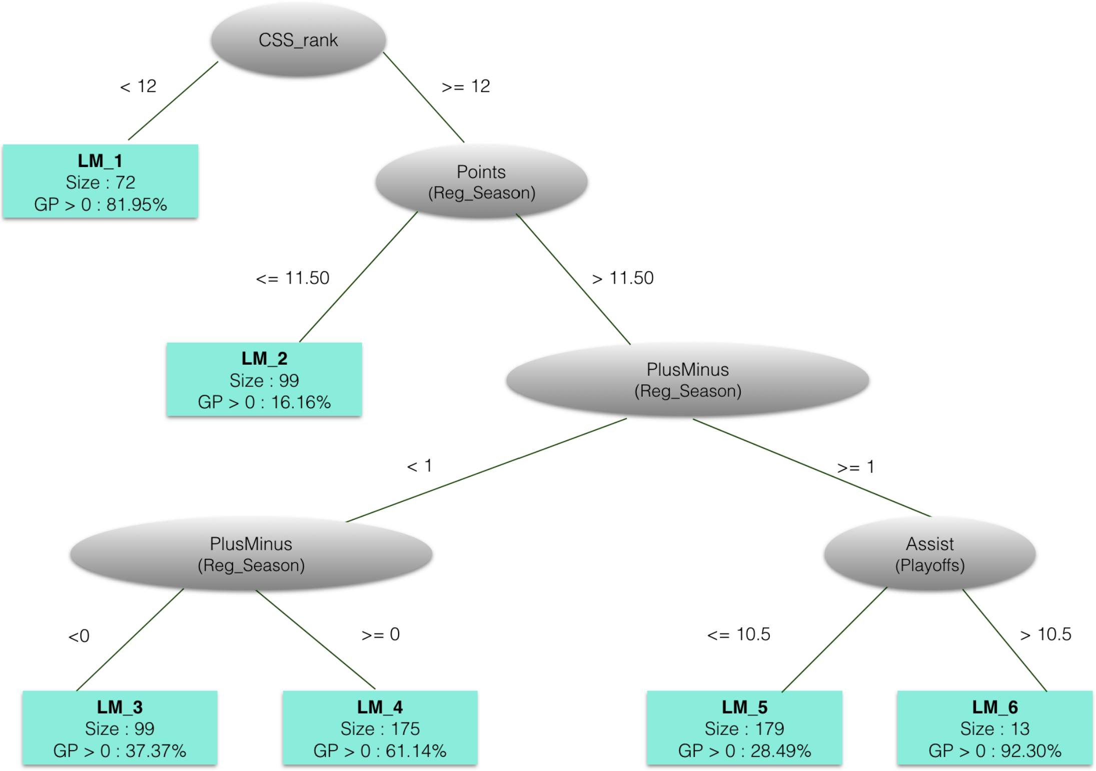

Figure 1. A logistic regression model tree for the 200(4+5+6) draft data. Each leaf defines a player group with its own predictive model. The most relevant root attribute is the Central Scouting Service rank. The tree assigns a single logistic regression equation to players that stand out with CSS rank higher than 12. The future success of lower-ranked players is more difficult to predict; they are partitioned into 5 groups, each with its own regression equation.

Group Definition	Group Equation	Top Player	Strongest Point
LM_1: CSS_rank <= 12	CSS_rank * -19.03 + po_PlusMinus * 4.56 + …	Patrick Kane	rs_Points = 154 vs. group mean = 41
LM_6: CSS_rank > 12 & Points > 11.5 & PlusMinus >= 1 & Assists > 10.5	(country = 'EURO') * -8.86 + po_GP * 1.08 + …	Brad Marchand	po_GP = 25 vs. po_GP_group = 19

Table 1. Illustrating the Two Most Successful Groups: Greatest-Magnitude Coefficients for Logistic Equation, Top Players, and their Strongest Features. Rs = Regular Season, po=Playoffs.

Conclusion

Model tree learning is easy to apply to draft data and produces a highly interpretable model ensemble. For each player, it can be used to predict future success, identify comparables, and highlight exceptional features. Model tree learning combines predictive modelling with descriptive analysis, in a manner that is both data-driven and intuitive for scouts, coaches, and other experts.

[1] Michael E. Schuckers (2016), 'Draft by Numbers: Using Data and Analytics to Improve National Hockey League Player Selection'. *MIT Sloan Sports Analytics*.

[2] Weissbock, J. (2015), 'Draft Analytics: Unveiling The Prospect Cohort Success Model'.