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

Msc. Thesis Defence
Yejia Liu

Outline

- Problem Formulation
- Previous Models
- Success Metrics
- Datasets
- Our Model Tree (NHL)
- Exceptional Players and Strongest Points Analysis
- NHL Case Studies
- Our Model Tree (NBA)
- NBA Case Studies
- Conclusions

Problem Formulation: Drafting Prospects

Data-driven Methods!

- Drafting: essential to build a successful team
  1. Scouts: expensive labour and hours
  2. Entry Draft (Lottery system): mistakes (e.g. "tanking games" issue)
     - ★ NHL: *Nikita Filatov* (6th) vs. *Erik Karlsson* (15th)
     - ★ NBA: *Sam Bowie* vs. *Michael Jackson (Portland Trail Blazers)*

Previous Models To Predict Player Performance

- Regression-based approaches
  - NHL
    - ⋆ GLM, ANN, SVM and LOESS by David Wison
    - ⋆ Generalized additive model by Schuckers using season-by-season data
    - ⋆ Markov model for play-by-play data
  - NBA
    - ⋆ Least square regression by Coates and Oguntimein
    - ⋆ Linear regression by Greene, using predraft + rookie years stats
- Similarity-based approaches: Prospect Cohort Sucess Model
  - ⋆ Prospect Cohort Sucess Model (scoring rate, height and age)
  - ⋆ PECOTA system in baseball
  - ⋆ Hierarchical clustering methods to cluster NBA players (Yale University)

Schuckers, M. & Statistical Sports Consulting, L. L. C. (2016), 'Draft by Numbers: Using Data and Analytics to Improve National Hockey League (NHL) Player Selection', *MIT Sloan Sports Analytics Conference.*

Greene, A. (2015), 'The Success of NBA Draft Picks: Can College Career Predict NBA Winners', Master's thesis, St.Cloud State University.

Success Metrics (Dependent Variable)

- NHL

- Games Played: for a player's first seven seasons
- Point Shares System (hockey-reference.com, season-by-season data)

  Skaters Point Shares = (marginal goals) / (marginal goals per point)

- ThoR by Schuckers and Curro: quantify the goal probability of a player action encompassing all on-ice events (play-by-play data)

- NBA

- Player Efficiency Rating (PER)
  - encompass both accomplishment and negative results of a player
  - aim to measure per-minute performance
  - average league PER is always 15, allowing to compare players across seasons
- Win Shares
  - Offensive Winshares: produced points, offensive possessions
  - Defensive Winshares: oppenent points, oppent possessions

Datasets and Independent Variables

Datasets Description

- NHL

Inputs	***junior league stats*** in the draft year (demographic + performance metrics)
Output	sum_7yr_GP
Training data	*cohort 1*: 1998, 1999, 2000 drafts; *cohort 2*: 2004, 2005, 2006 drafts
Testing data	*cohort 1*: 2001, 2002 drafts; *cohort 2*: 2007, 2008 drafts

- NBA

Inputs	***college stats*** in the draft year (demographic + performance metrics)
Output	PER (player efficiency rating)
Training data	1985-2005 drafts, inclusive
Testing data	2006-2011 drafts, inclusive

Datasets

Datasets Preprocessing

- NHL

- Aggregate last season performance stats across teams
- Replacing missing values (missing CSS_rank = maximum rank + 1)
- Excluding players drafted in 2003

- NBA

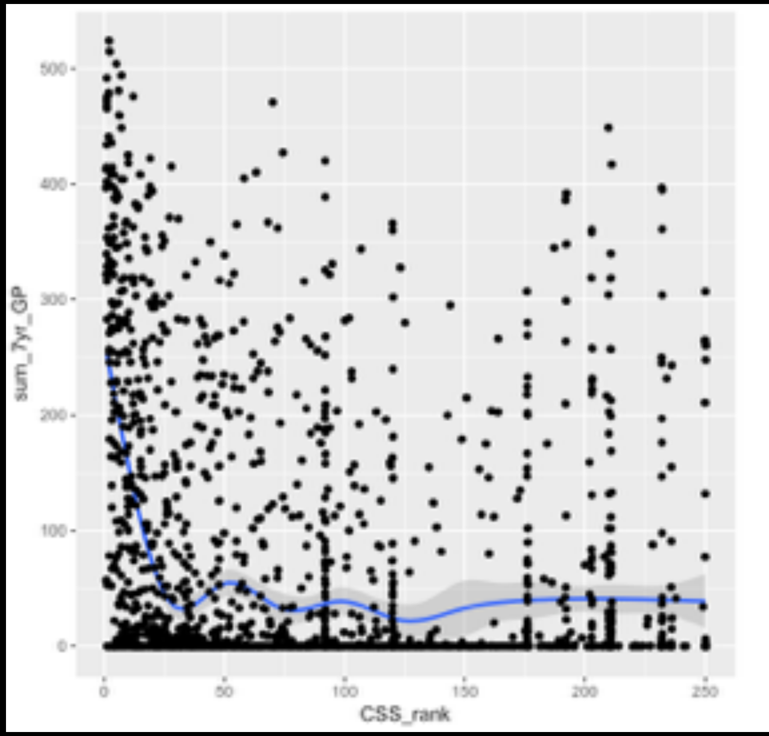
college stats	NBA stats	count	preprocessing
1	0	15	replaced NBA stats by min(x)-std(x)
0	1	173	excluded
0	0	35	excluded
1	1	1405	kept

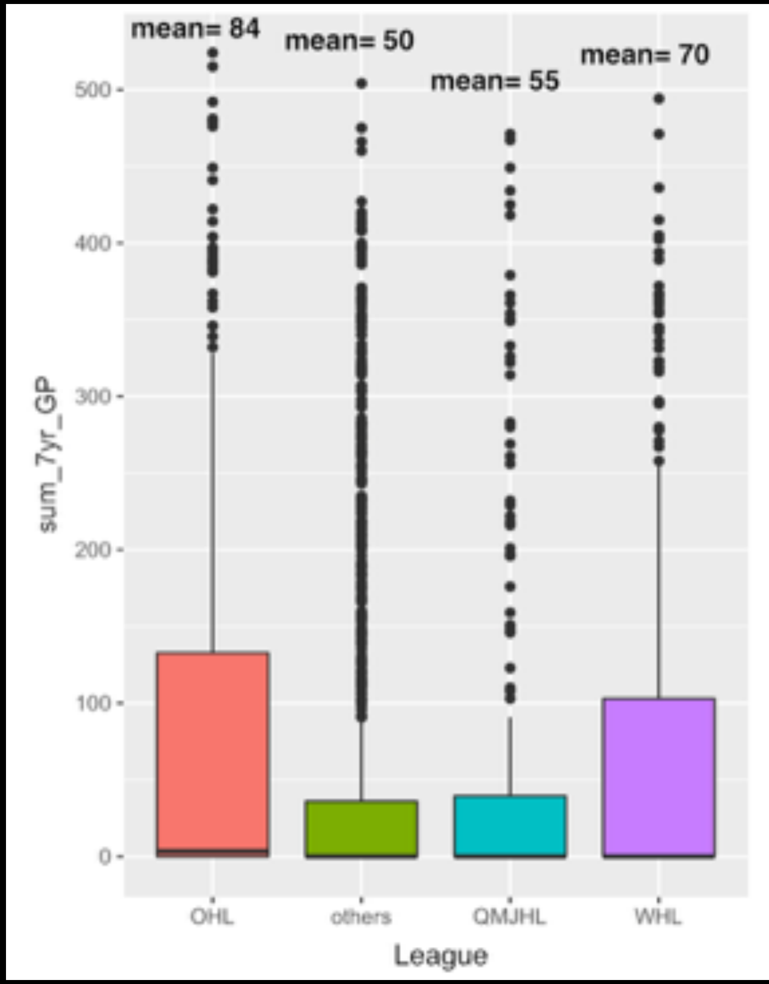
Datasets

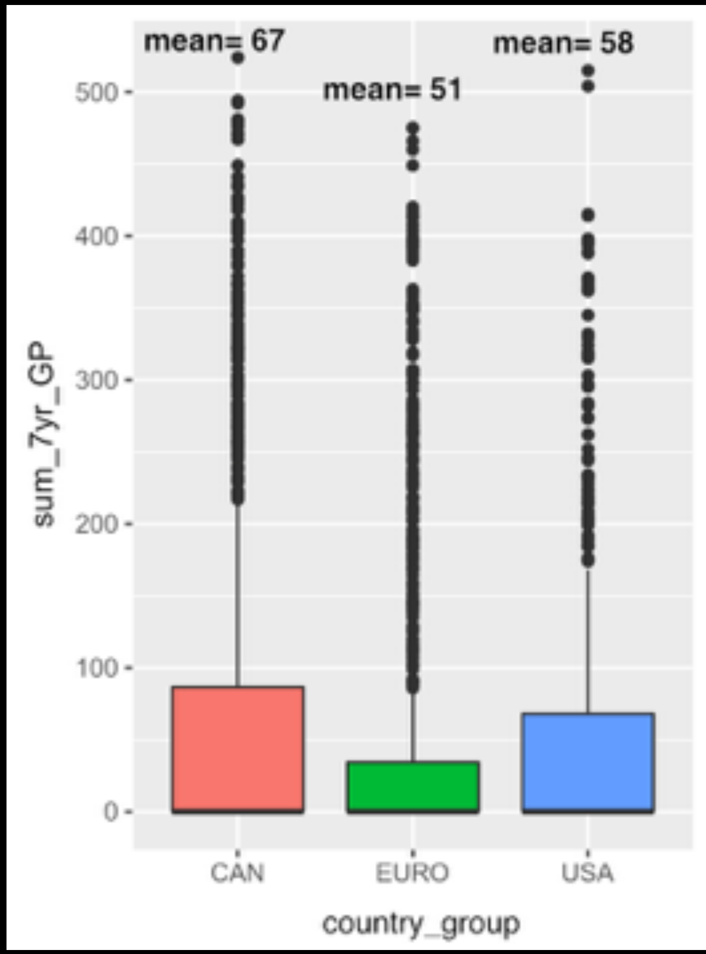
Datasets Exploration

- NHL (wrt sum_7yr_GP)

- CSS_rank (ranking from scouts)
- Major junior league
- country_group







Datasets

Datasets Exploration

- NBA (wrt PER)

- position (sorted by *mean*, descendant)

position	size	mean	std	min	25%	50%	75%	max
Center and Power Forward	162	14.93	3.5	7	11.85	13.75	16.33	24.6
Shooting Guard and Point Guard	115	13	3.28	4.4	10.7	12.65	14.75	24.2
Power Forward and Small Forward	108	12.94	3.35	-1.5	11.1	13.35	15.9	20.8
Small Forward and Shooting Guard	68	12.81	3.03	7	10.68	12.65	14.45	25.2
Point Guard	181	12.61	6.7	-6.8	9.7	12.2	15.1	76.1
Power Forward	134	11.95	6.94	-30.2	9.8	11.85	14.78	58.3
Small Forward	142	11.05	4.77	-5.6	8.7	11.05	13.9	31.3
Shooting Guard	142	10.66	4.9	-11.4	8.58	11.45	13.4	22.2
Center	227	9.96	9.78	-48.6	8.6	11.3	13.8	66.8
Guard	10	-14.2	18.97	-57.62	-24.25	-7.87	-1.07	1.61
Forward	12	-14.87	20.73	-57.62	-15.13	-5.48	-1.44	-0.88
Forward/Center	7	-14.24	13.58	-27.62	-27.62	-15.13	-1.63	1.61

Our Model Tree

- Combine regression-based and similarity-based approaches
    - ★ An ensemble of regression models
    - ★ Present interactions between player features and player groups
    - ★ Learn from data, no need to specify similarity metrics
    - ★ Differentiate players from the same group

Our Model Tree (NHL)

- How we build the model tree
  - Zero-inflation problem in NHL draft (about half of player not playing in NHL after being drafted)
  - *Whether a drafted player can play at least one game at NHL?*
  - Logistic regression model in the leaf node
  - Process:

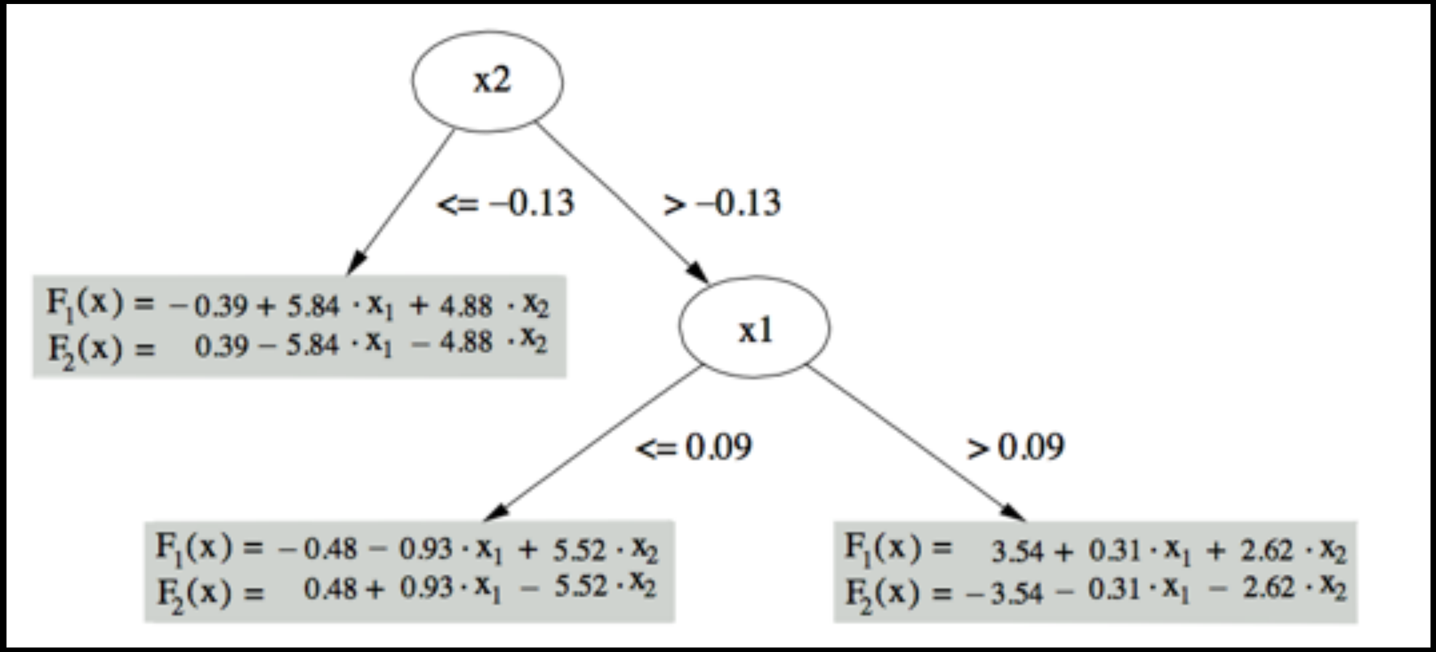
    1. Build a tree whose leaves contain a logistic regression model.

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

    3. Use $m(l_i)$ to compute a probability $p_i = P(g_i > 0)$.

Our Model Tree (NHL)

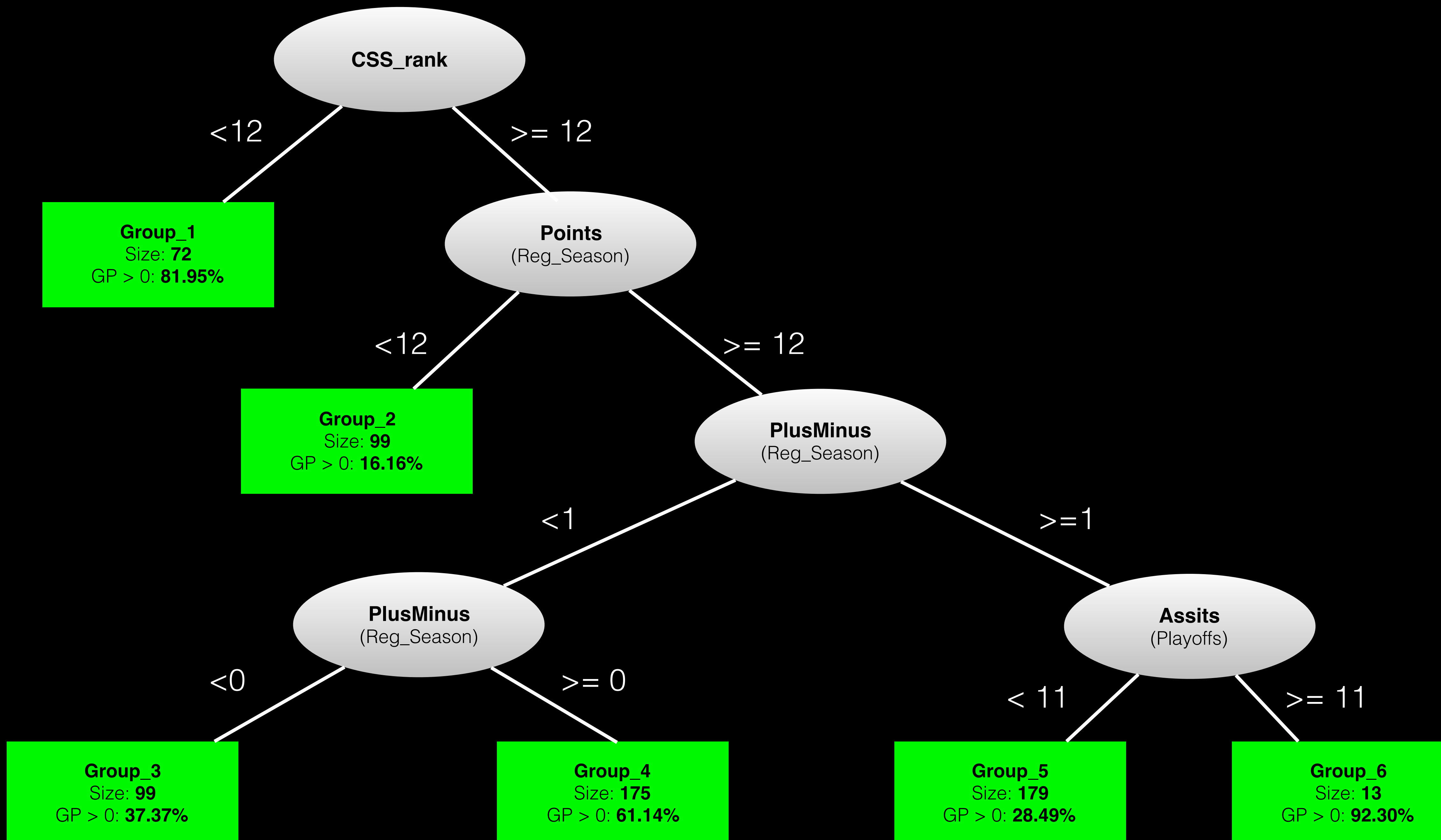
- Logistic Model Trees
  - ⋆ Logistic regression model in every node
  - ⋆ LogitBoost algorithm to maximize likelihood of training data points



* Example Tree

  - ⋆ Tree splitting based on information entropy, similar to C4.5
  - ⋆ Tree pruning based on training error and model complexity penalty

Our Model Tree (NHL)



Model tree for 2004, 2005, 2006 cohort of drafted NHL players

Our Model Tree (NHL)

- Evaluation (Spearman Rank Correlation)

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**
2001, 2002	2002	0.3	**85.79%**	**0.85**
2004, 2005, 2006	2007	0.46	**81.23%**	**0.84**
2007, 2008	2008	0.51	**63.56%**	**0.71**

Exceptional Players and Strongest Points

Calculation Methods

- We can leverage the weights to identify the player features that contribute the most to raising/lowering a player's ranking
- The probability difference of playing at least one game between a random player *i* and an average player in group *g is:*

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

- Find the features *j* that contribute the most to this difference:

$$argmax_j \; |w_j(x_{ij} - \overline{x_{gj}})|$$

NHL Case Studies

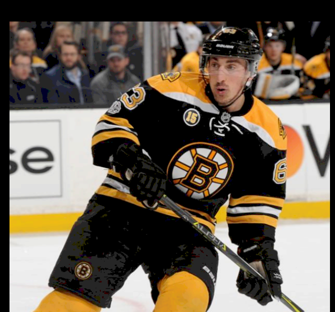
- Underestimated Player: Kyle Cumiskey, Brad Marchand

6	Brad Marchand	Country	po_GP	po_P
		CAN	25 ($\bar{x}$ = 19)	23 ($\bar{x}$ = 19)
	Mathieu Carle	Country	CSS_rank	rs_GP
		CAN	53 ($\bar{x}$ = 107)	67 ($\bar{x}$ = 65)
	Kyle Cumiskey	Country	po_GP	rs_GP
		CAN	27 ($\bar{x}$ = 19)	72 ($\bar{x}$ = 65)

- Not ranked by CSS at all
- Overall pick at 222

His strongest points are identified as *GP*

★ 132 NHL games, Won a Stanley Cup (2015), Represented Canada in the World Championship



- Ranked 80 by CSS
- Overall pick at 71

His strongest points are identified as *playoff* stats

★ 534 NHL games, won a Canada Cup/World Cup (2016)

Our Model Tree (NBA)

- How we build the model tree

  - No zero-inflation problem in NBA draft, over 80% drafted player played in NBA
  - Predict career PER of a drafed player
  - Process:

    1. Build a tree whose leaves contain a linear regression model.
    2. The tree assigns each player i to a unique leaf node l_i, with a linear regression model $m(l_i)$.
    3. Use $m(l_i)$ to compute predicted career PER.

Our Model Tree (NBA)

- M5 Regression Trees (M5P)

- Intial tree construction based on standard deviaton of target variables

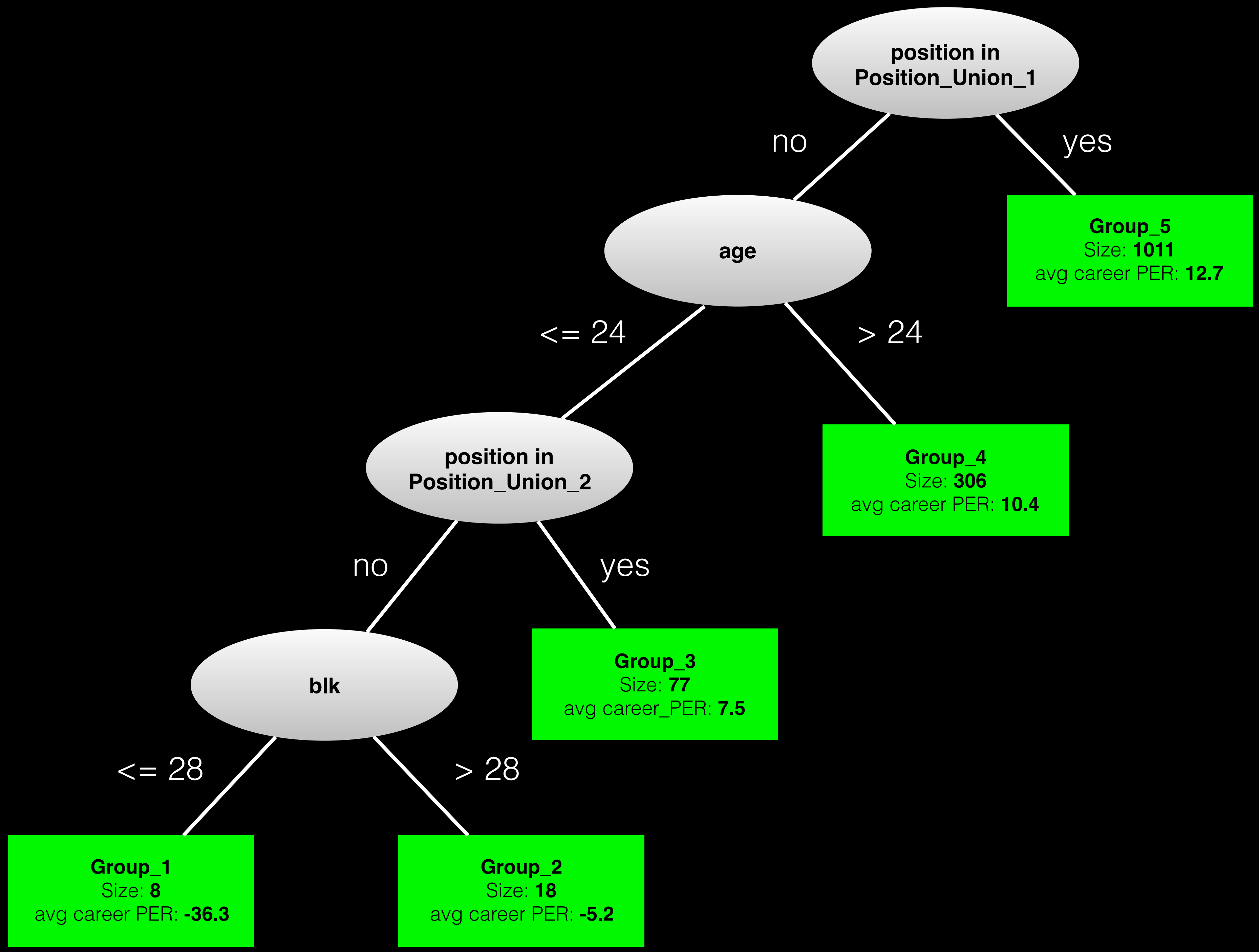
$$\Delta error = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

- Linear regression model in every node using standard regression methods

- Tree pruning based on estimated error

- Tree Smoothing: predicted value at leaf node adjusted by the predicted values from root to this leaf node

Our Model Tree (NBA)



Model tree for players drafted between 1985-2011 in NBA

Our Model Tree (NBA)

- Evaluation

	Pearson Correlation	Spearman Rank Correlation	RMSE
Draft Order	0.42	0.39	NaN
Linear Regression(baseline)	0.45	0.40	7.14
Our Model Tress	**0.55**	**0.43**	**6.16**

NBA Case Studies

- Underestimated Player: Dejuan Blair

draft year	draft pick	career PER	predicted PER	comparables(career_per, draft pick)
2009	37	16.5	17.2	Jordan Hill (16.3, 8th)

Conclusion

- Introduce model trees, which
  - assign players to groups that are statistically distinct
  - build separate prediction models for separate groups
- Model tree rankings correlate with actual career success metric (sum_7yr_GP, career PER)
- Tree structure is interpretable for scouts, sport experts
- Model trees can be used to highlight player strong/weak points
- Our methods are flexible to apply to other sports with aggregate datasets

Thank you!!!

&

Questions?