NHL PLAYOFF MODEL

Ethan Greig
2018-04-10

PURPOSE

- Regression model for NHL playoff series winners
- Direct comparison of opponents' team statistics
- Teams are not given individual scores

	Wins	Goals For	Goals Against
Team A	42	254	241
Team B	36	279	278
Difference	+6	-25	-37

Team A Win Probability: 0.57

DATA COLLECTION

- 10 seasons of regular season and playoff data (since 2007-08)
- Sourced from various websites, by exporting or copy-pasting to CSV
  - NHL.com
  - Corsica Hockey
  - MoreHockeyStats.com
- Combined into one R dataframe

INPUT VARIABLES

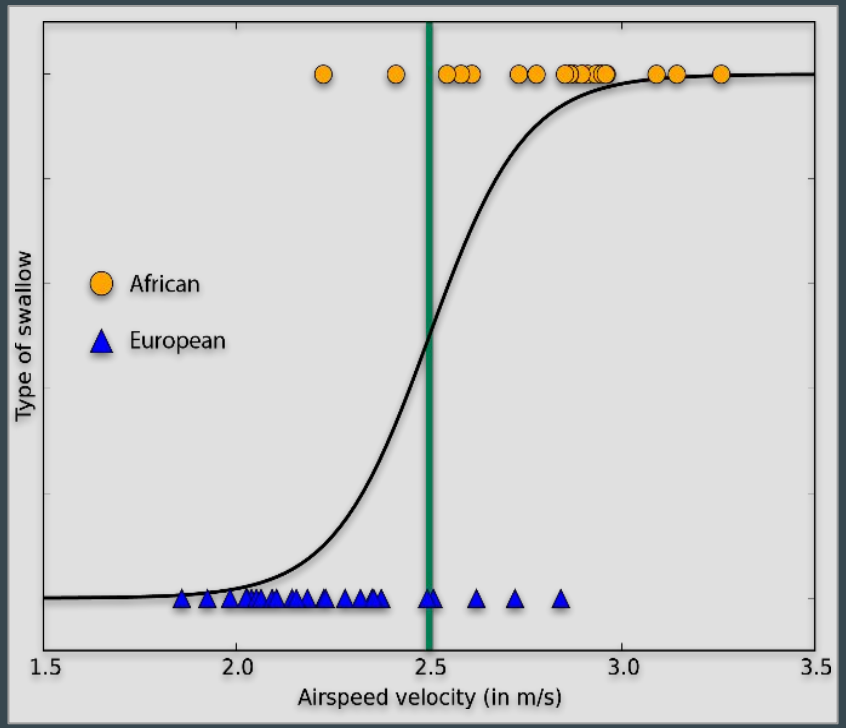
GF%	Goals For Percentage
AdjGF%	Non-Empty Net Goals For Percentage
5v5_xGF%	5-on-5 Expected Goals For Percentage
5v5_CF%	5-on-5 Shot Attempts For Percentage
5v5_Sh%	5-on-5 Team Shooting Percentage
5v5_Sv%	5-on-5 Team Save Percentage
PP%	Power Play Percentage
PK%	Penalty Kill Percentage
Time_Led	Minutes Per Game With the Lead
ROWins	Number of Non-Shootout Wins

VARIABLE CORRELATIONS

- Calculate correlation between each team statistic and team playoff success
- Playoff success is measured by number of playoff wins

	GF%	Adj GF%	5v5 xGF%	5v5 CF%	5v5 Sh%	5v5 Sv%	PP%	PK%	Time Led	R+OT Wins
5 years	0.35	0.34	0.24	0.23	-0.07	0.17	-0.10	0.22	0.07	0.12
10 years	0.34	0.33	0.21	0.18	-0.03	0.11	0.06	0.20	0.21	0.15

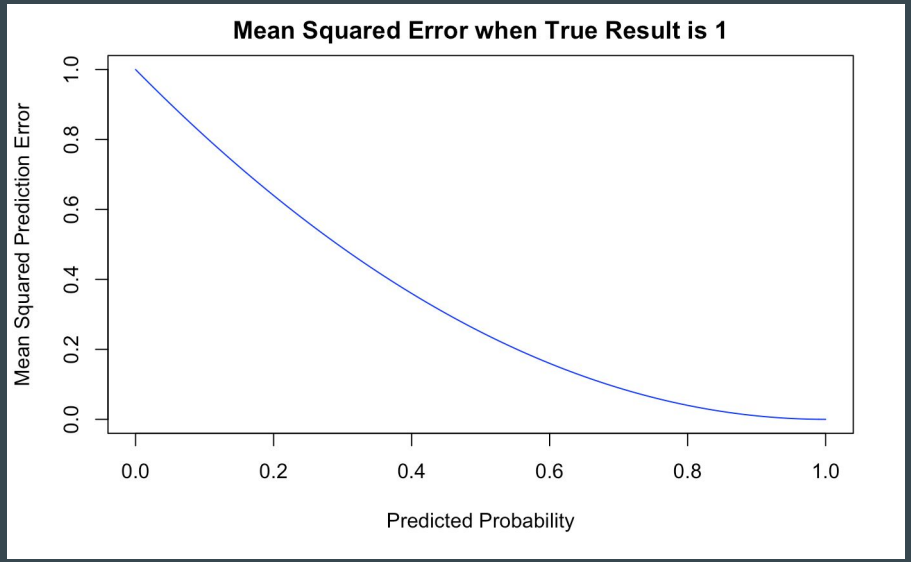
LOGISTIC REGRESSION



- Selected a subset of the variables as input variables for the model
- Difference between higher and lower seeded team is calculated for each statistic
- Used playoff series result as response variable
  - Binary (won/lost)
  - Weighted (win% in series)

MODEL VALIDATION

- Iterated 10-fold cross validation
- Loss function averaged over test sets
  - Log Loss (~0.56)
  - Mean Squared Prediction Error (~0.23)
- Resulted in some unintuitive coefficients



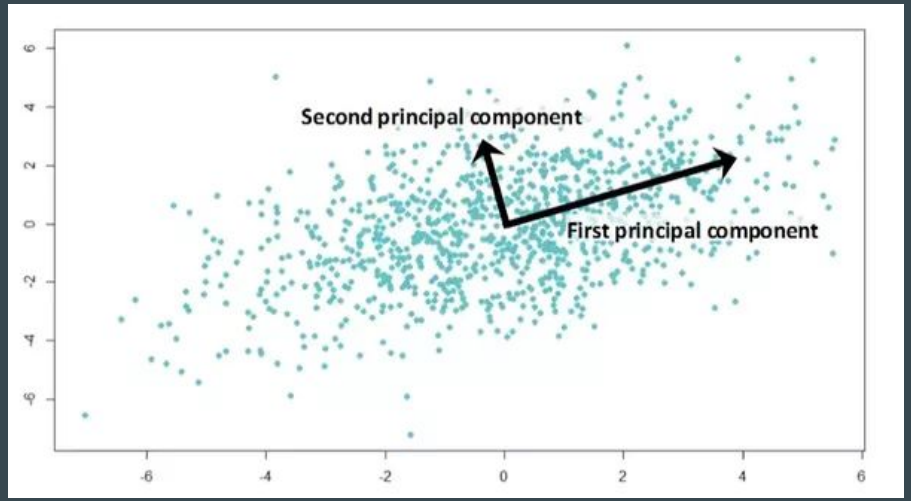
MULTICOLLINEARITY

	GF%	Adj GF%	5v5 xGF%	5v5 CF%	Time Led	5v5 Sh%	5v5 Sv%	PP%	PK%	R+O Wins
GF%	1.0	1.0	0.3	0.3	0.6	0.4	0.3	0.3	0.2	0.5
Adj GF%	1.0	1.0	0.3	0.3	0.6	0.4	0.3	0.3	0.2	0.4
5v5 xGF%	0.3	0.3	1.0	0.8	0.2	-0.2	-0.2	0.0	0.1	0.1
5v5 CF%	0.3	0.3	0.8	1.0	0.1	-0.4	-0.2	0.0	0.1	0.1
Time Led	0.6	0.6	0.2	0.1	1.0	0.4	0.1	0.3	0.0	0.3
5v5 Sh%	0.4	0.4	-0.2	-0.4	0.4	1.0	-0.1	0.1	-0.1	0.2
5v5 Sv%	0.3	0.3	-0.2	-0.2	0.1	-0.1	1.0	0.0	-0.1	0.1
PP%	0.3	0.3	0.0	0.0	0.3	0.1	0.0	1.0	-0.2	0.2
PK%	0.2	0.2	0.1	0.1	0.0	-0.1	-0.1	-0.2	1.0	0.0
R+O Wins	0.5	0.4	0.1	0.1	0.3	0.2	0.1	0.2	0.0	1.0

- Many of the 10 input variables are highly correlated
- Statistical variation is hidden in complicated multicollinearity
- Regression models work much better with independent variables

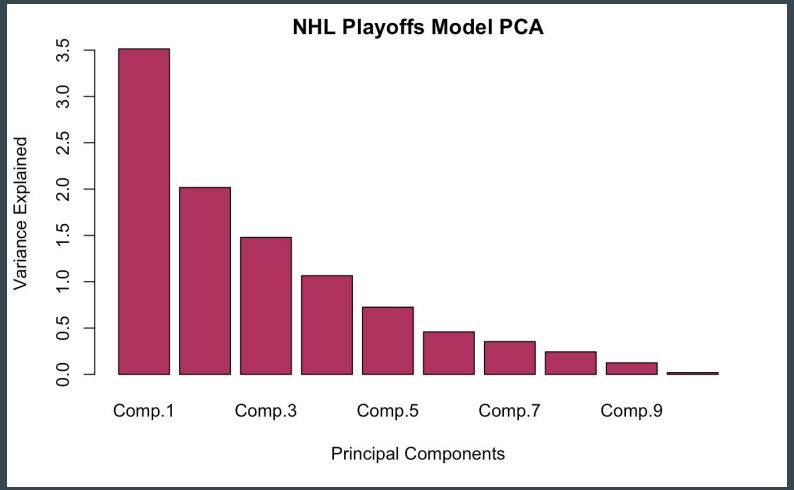
PRINCIPAL COMPONENT ANALYSIS

- Extracts principal components from (normalized) input variables
- These linear combinations of the variables are uncorrelated
- Resulting components are sorted by decreasing impact on total variation
- Whiteboard example of PCA



FINAL MODEL

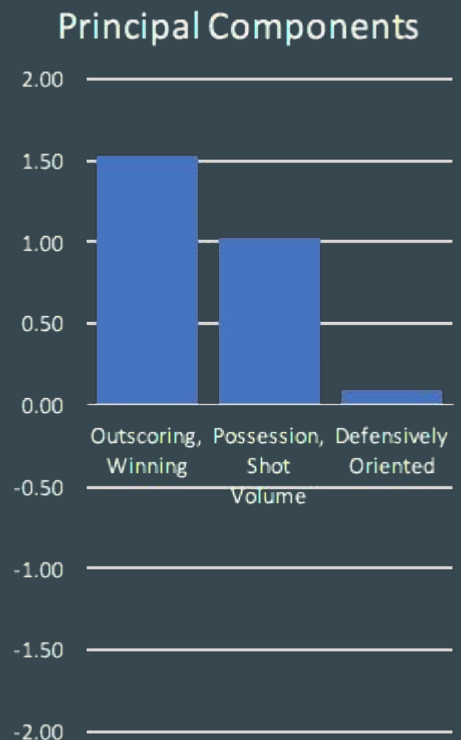
- Three components explain 70.1% of variance
  - Component 1: “Outscoring and Winning”
  - Component 2: “Possession and Shot Volume”
  - Component 3: “Defensive Inclination”
- Other components do not have clear intuition



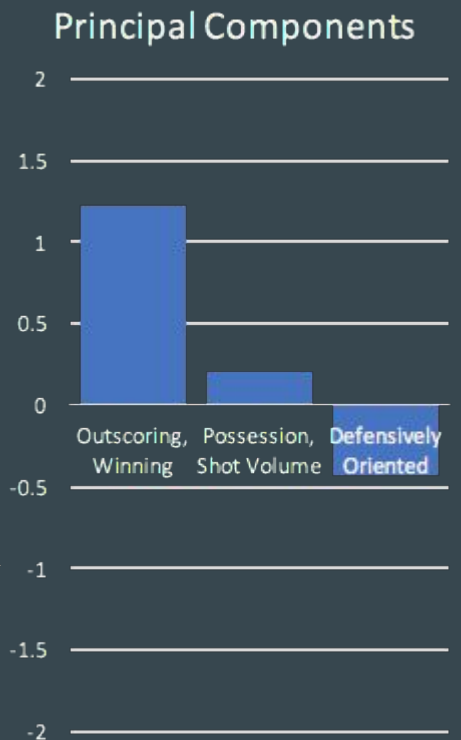
	Comp.1	Comp.2	Comp.3
ΔGF%	0.49	-0.09	0.24
Δ Adj GF%	0.48	-0.09	0.24
Δ 5v5 xGF%	0.27	0.50	-0.14
Δ 5v5 CF%	0.24	0.57	-0.06
Δ Time Led	0.41	-0.05	-0.18
Δ 5v5 Sh%	0.14	-0.53	-0.13
Δ 5v5 Sv%	-0.04	-0.19	0.58
Δ PP%	0.17	-0.16	-0.50
Δ PK%	0.08	0.17	0.48
Δ R+O Wins	0.41	-0.20	-0.07

PREDICTIONS (CENTRAL)

75% ↑

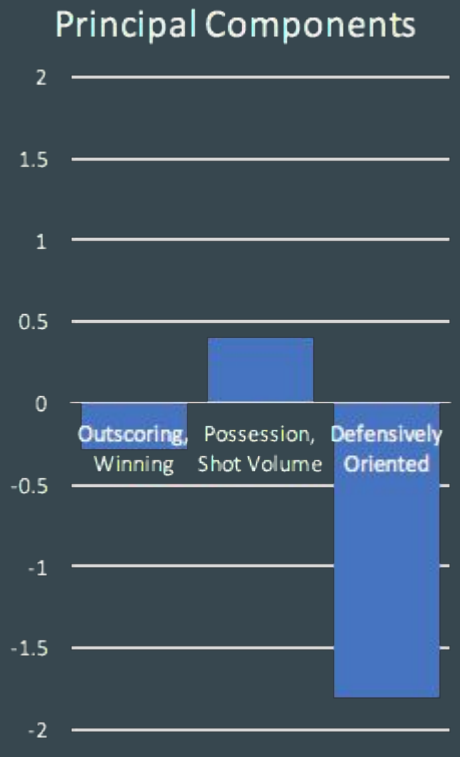


64% ↑

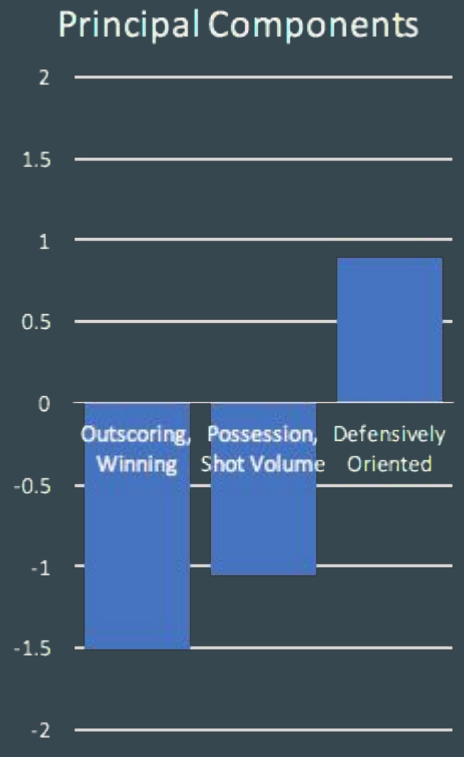


PREDICTIONS (PACIFIC)

36% ↓

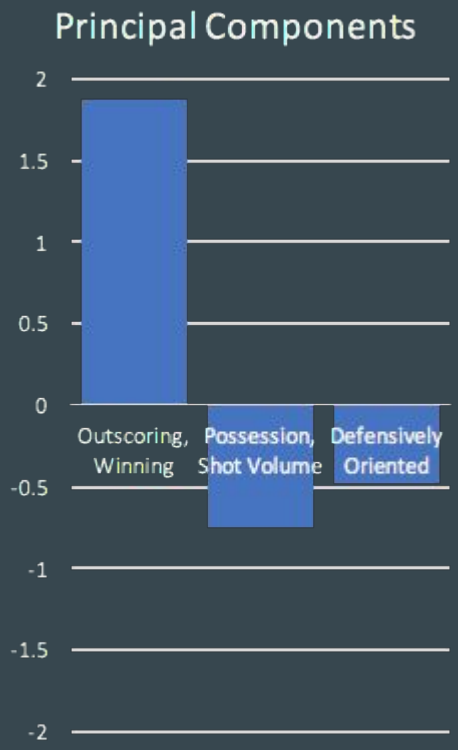


47% ↓

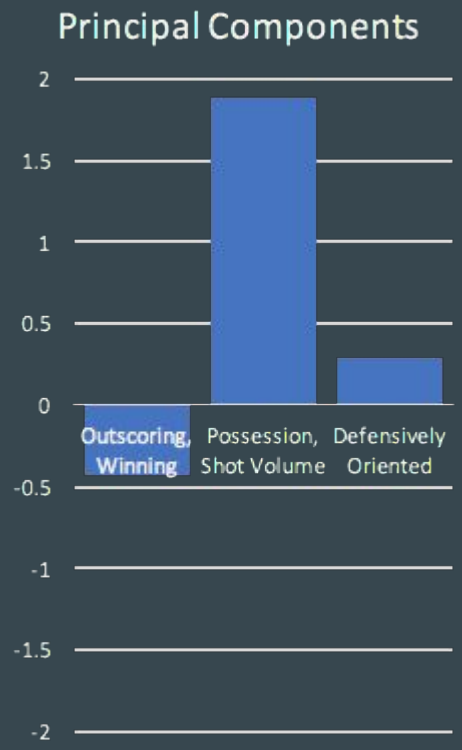


PREDICTIONS (ATLANTIC)

64% ↑

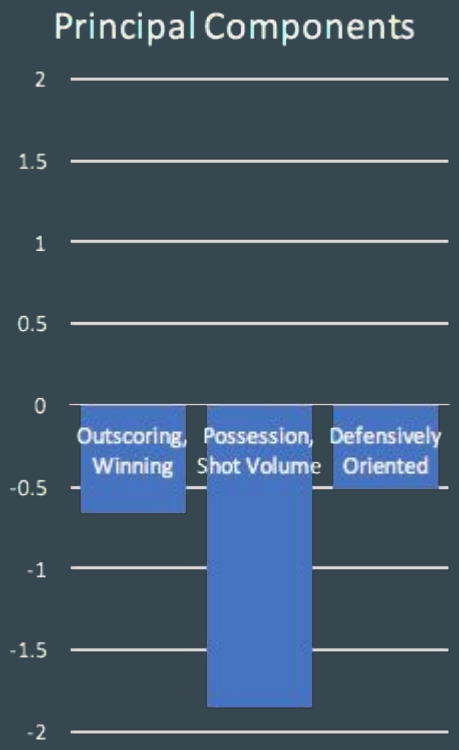


65% ↑



PREDICTIONS (METROPOLITAN)

36% ↓



50% ↑