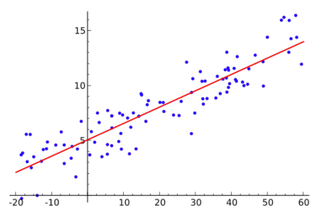


NHL Predictive Analytics

BACKGROUND

The goal of this project was to determine the most influential statistics in predicting the regular season success of NHL teams and also predict the outcome of individual games. All data used is from SAP and Peter Tanner at moneypuck.com.

GOALS

1) Predict number of regular season ROW (Regulation or Overtime Wins) each NHL team accrued based on their average statistics through all 82 games.
2) Predict the outcome of NHL regular season games based on each team's average statistics in their 15 prior games.

METHODS

Use cross validation with training and testing sets of data.

1) Linear and Binomial Regression via both Backwards Elimination and AIC.
2) Logistic Regression, Decision Tree, and Random Forest classification techniques.

PREDICTING NUMBER OF WINS:

Regression	Model Selection	MSE
Linear	Backward Elimination	5.35
	AIC	5.35
Binomial	Backward Elimination	5.84
	AIC	5.84

BEST MODEL (Linear Model with 6 predictors):

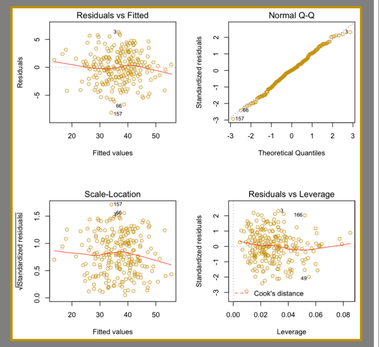
- Goals For, PIM For, Shots Against, Goals Against, Rebounds Against, Hits Against

$$Y = 40.47 + 16.14X_{GF} - 0.36X_{PIM} - 0.29X_{SOG.Ag} - 13.60X_{GA} + 2.19X_{RB.Ag} - 0.15X_{Hits.Ag}$$

PREDICTIONS

Team	Season	Number.of.Wins	Predicted.Wins
FLA	2017	41	38
DET	2014	39	40
CAR	2010	35	37
FLA	2008	38	38
NYR	2009	35	35
T.B	2013	38	41

DIAGNOSTICS



CONCLUSIONS:

- Linear model predicts more accurately than binomial or Poisson model.
- Defense is arguably more important than offense in terms of regular season success.

PREDICTING SINGLE GAME OUTCOMES:

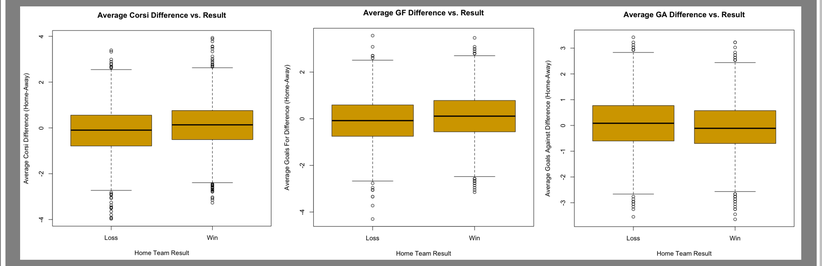
Classification Method	Accuracy
Logistic Regression	57.68%
Random Forest	54.03%
Support Vector Machines	56.62%

BEST MODEL (Logistic Regression with 3 predictors):

- Corsi Percentage Difference, Goals For Difference, Goals Against Difference

$$logit(P(Win)) = -0.054 + 0.19X_{Corsi.d} + 0.16X_{GF.d} - 0.15X_{GA.d}$$

PREDICTORS BY WINS & LOSSES



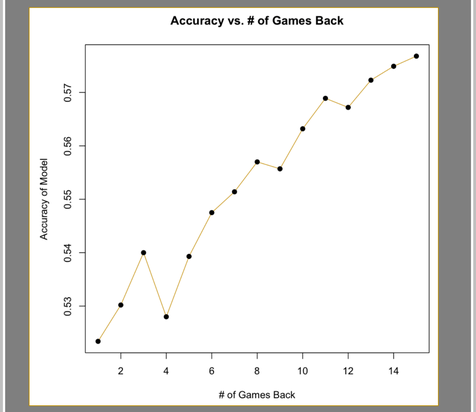
EXAMPLE PREDICTIONS

Home.Team	Away.Team	Corsi.Diff	Goals.Against.Diff	Goals.For.Diff	Home.Result	Home.Pred.Results
FLA	WPG	-0.5883315	-2.07414558	-1.7113548	Loss	Loss
WPG	COL	-0.1365746	0.67303700	2.1245197	Win	Win
S.J	DET	0.3855535	0.47680967	-0.1770050	Loss	Loss
NYR	CHI	-0.5433762	0.08435502	1.3573448	Loss	Win
BOS	CBJ	0.6358929	-1.48546360	0.3983762	Win	Win
MIN	BOS	-0.8610017	-0.70055429	-2.6703235	Loss	Loss

CONFUSION MATRIX

n=1503	Predicted: Loss	Predicted: Win
Actual: Loss	481	363
Actual: Win	273	386

ACCURACY VS. GAME HISTORY



CONCLUSIONS:

- Logistic regression model achieved higher accuracy than RF or SVM with the predictors used.
- Model predicts slightly less accurate than Vegas, which predicts NHL games correctly about 60% of the time.

COMPLICATIONS/LESSONS:

- Cleaning and preparing the data for analyses took much longer than anticipated.
- Properly cleaning the data saves time in the end.

FUTURE WORKS:

- Use more advanced statistics and perhaps calculate my own advanced stats to achieve higher accuracy.
- Compare teams via ratios rather than differences.

Special Thanks to: