MTH 522 Presentation: Predicting NHL Shots on Goal

Brian Cornet

April 27th, 2021

Background

- The **National Hockey League (NHL)** provides data publicly from its API:
  - https://statsapi.web.nhl.com/api/v1
- This dataset includes all game "plays" from 2000-2001 season to 2019-2020 season:
  - https://www.kaggle.com/martinellis/nhl-game-data
- Has been suggested that predictive modeling for hockey is extremely difficult compared to other sports

Shot Blocked:
The puck is stopped or redirected by a skater using either their stick or their body (ouch)

Shot on Goal:
The puck is stopped or redirected by the goalie OR the puck goes in the net (goal!)

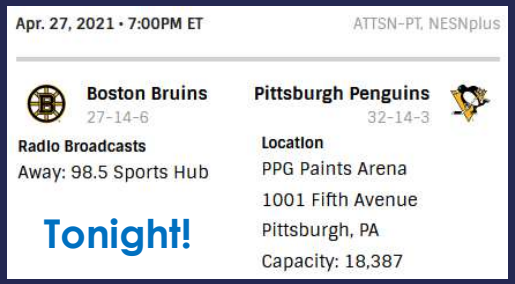
Data Overview

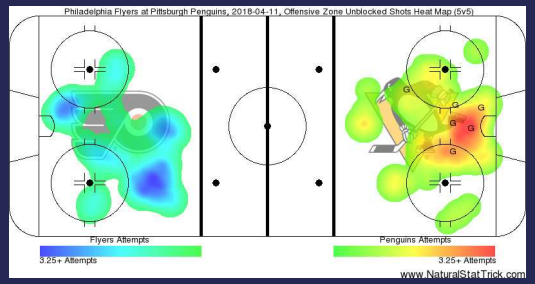
This model attempts to predict **shots on goal** vs. **shots blocked by skaters** (as reported by the NHL API) using API-provided variables:

- **play_id, game_id:** index values for individual plays (game_plays.csv) and games (game.csv)
- **team_id_for, team_id_against:** index values for involved players (team_info.csv)
- **event:** category for a play; Faceoff, Hit, Penalty, Shot, Goal, etc.
- **secondaryType:** subcategory for certain events such as shot types (slap shot, deflection, etc.)
- **x,y, st_x,st_y:** coordinates on ice (in feet); absolute and relative to player's goal respectively
- **period, periodType:** REGULAR for periods 1-3, OVERTIME for 4+, SHOOTOUT for 5 (regular season)
- **periodTime, periodTimeRemaining:** time (in seconds) since start and until end of period
- **dateTime:** timestamp of event (GDT)
- **goals_away, goals_home:** current score based on home/away teams (NOT team_id_for/against)
- **description:** full text description of event; players involved, actions, etc.

Attributes Used

- Joined dataset with game.csv to make the following changes:
  - Added **playoffs**; 0 = regular season, 1 = Stanley Cup Playoffs (generally more intense)
  - Added **season** as integer for season number (e.g. 2019-2020 → 20)
  - Identified shooter by home/away status to create **shooterHome**; 0 = away team, 1 = home team
  - Mapped **goals_home** and **goals_away** to **shooterGoals** and **defenderGoals** based on the above
- Binarized **periodType** to **overtime** and **shootout**; **event** to **onGoal** (1 = shot/goal, 0 = block)
- **12 predictors used** that describe the shot setting, shot time, and shot position





Sample Selection

- Focus is on shots blocked vs. shots on goal: **event** must be "Blocked Shot", "Shot", "Goal"
- Removed all entries with NA values in relevant attributes using *is.na()* from R base
  - Includes all entries prior to 2010-2011 season; this is when NHL's API started reporting coordinates
- Removed all duplicate entries in dataset using *duplicated()* from R base
  - All entries in Kaggle dataset starting from the 2018-2019 season are listed twice for some reason
- Split dataset into two without replacement using *createFolds()* from R **caret** package
- **Final data size:** 1,054,130 entries **Training set size:** 527,065 **Testing set size:** 527,065

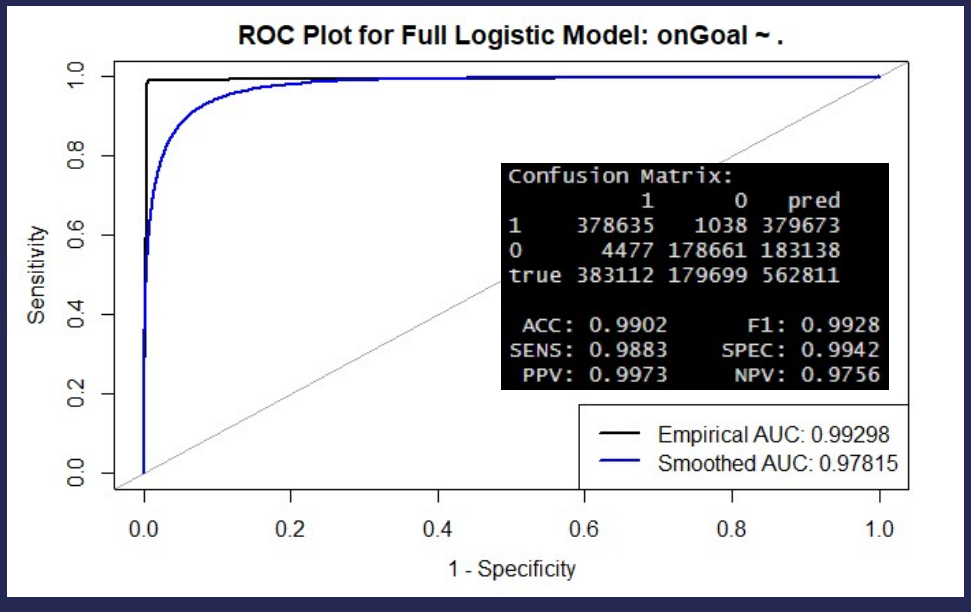
```
     season         playoffs         shooterHome       shooterGoals     defenderGoals        period
 Min.   :11.00   Min.   :0.00000   Min.   :0.0000   Min.   : 0.000   Min.   : 0.000   Min.   :1.000
 1st Qu.:13.00   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:1.000
 Median :16.00   Median :0.00000   Median :1.0000   Median : 1.000   Median : 1.000   Median :2.000
 Mean   :15.63   Mean   :0.07614   Mean   :0.5057   Mean   : 1.335   Mean   : 1.286   Mean   :2.048
 3rd Qu.:18.00   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.:3.000
 Max.   :20.00   Max.   :1.00000   Max.   :1.0000   Max.   :10.000   Max.   :10.000   Max.   :8.000
    overtime          shootout          periodTime      periodTimeRemaining       st_x              st_y
 Min.   :0.00000   Min.   :0.000000   Min.   :   0.0   Min.   :   0.0      Min.   :-99.00   Min.   :-42.00000
 1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.: 286.0   1st Qu.: 284.0      1st Qu.:-55.00   1st Qu.:-12.00000
 Median :0.00000   Median :0.000000   Median : 586.0   Median : 589.0      Median : 48.00   Median :  0.00000
 Mean   :0.01808   Mean   :0.006487   Mean   : 593.7   Mean   : 587.4      Mean   : 19.96   Mean   :  0.00883
 3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.: 897.0   3rd Qu.: 890.0      3rd Qu.: 71.00   3rd Qu.: 12.00000
 Max.   :1.00000   Max.   :1.000000   Max.   :1200.0   Max.   :1200.0      Max.   : 99.00   Max.   : 42.00000
```

```
     onGoal
 Min.   :0.0000
 1st Qu.:0.0000
 Median :1.0000
 Mean   :0.6812
 3rd Qu.:1.0000
 Max.   :1.0000
```

onGoal	n
0	358832
1	766791

Logistic Regression Model

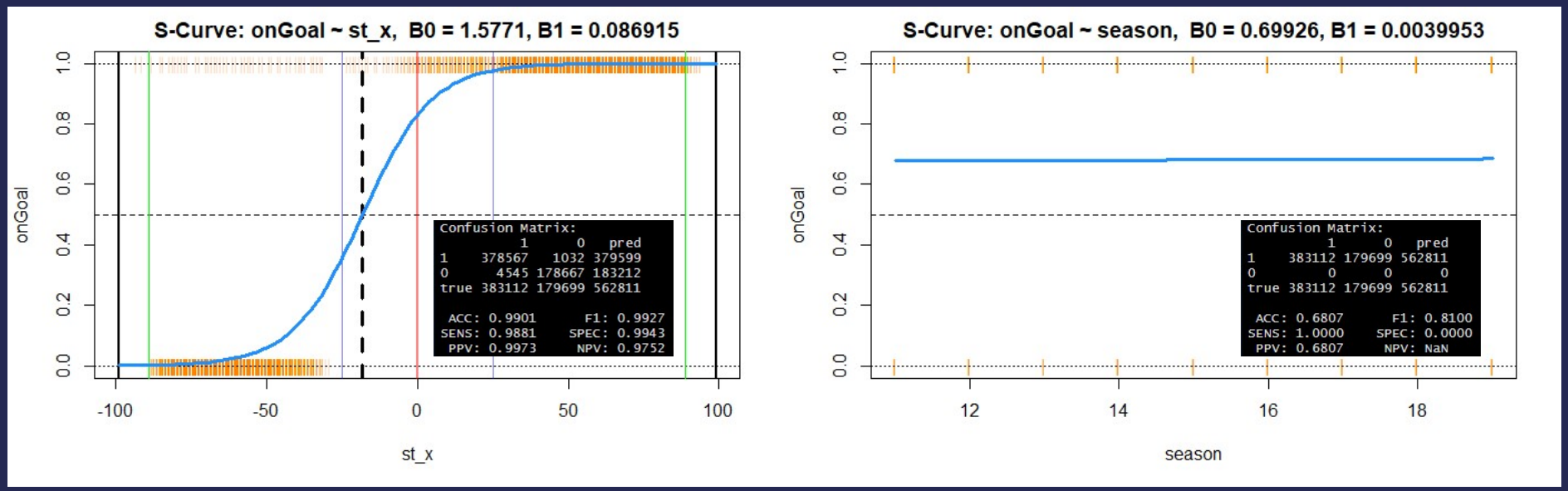
- Very high accuracy for this model!
  - **st_x** z-value of 258, p-value floored to 0 in memory
  - **season** z-value of 4.527, p-value of 2.122916e-113



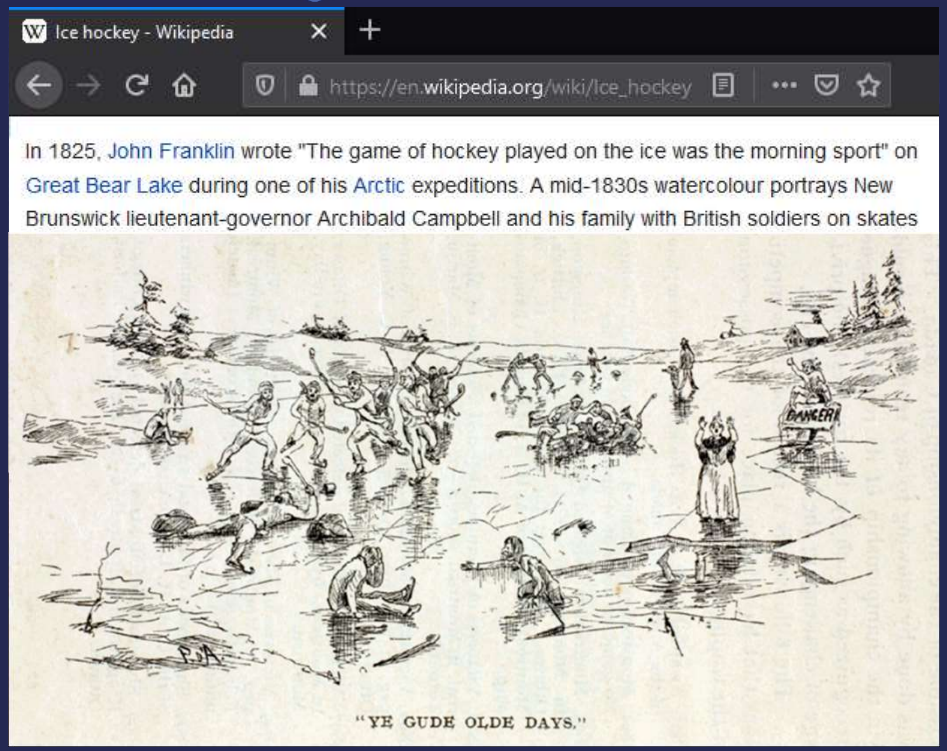

```
Call:
glm(formula = data.formula, family = binomial(link = "logit"),
    data = data.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6038  -0.1128   0.0228   0.0578   3.8548

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          2.309e-02  3.210e-01   0.072   0.9426
season               9.782e-02  4.322e-03  22.631  < 2e-16 ***
playoffs             8.185e-02  4.522e-02   1.810   0.0703 .
shooterHome          2.468e-02  2.422e-02   1.019   0.3082
shooterGoals         2.000e-02  1.067e-02   1.874   0.0609 .
defenderGoals        4.961e-02  1.096e-02   4.527 5.99e-06 ***
period               3.782e-02  1.996e-02   1.894   0.0582 .
overtime            -3.334e-01  1.995e-01  -1.671   0.0947 .
shootout             1.853e+01  8.930e+01   0.208   0.8356
periodTime          -5.347e-05  2.628e-04  -0.204   0.8387
periodTimeRemaining -2.055e-04  2.622e-04  -0.784   0.4330
st_x                 8.781e-02  3.401e-04 258.205  < 2e-16 ***
st_y                 1.172e-03  6.352e-04   1.845   0.0650 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 704154  on 562811  degrees of freedom
Residual deviance:  59890  on 562799  degrees of freedom
AIC: 59916

Number of Fisher Scoring iterations: 9
```

Simple Logistic Regression Models

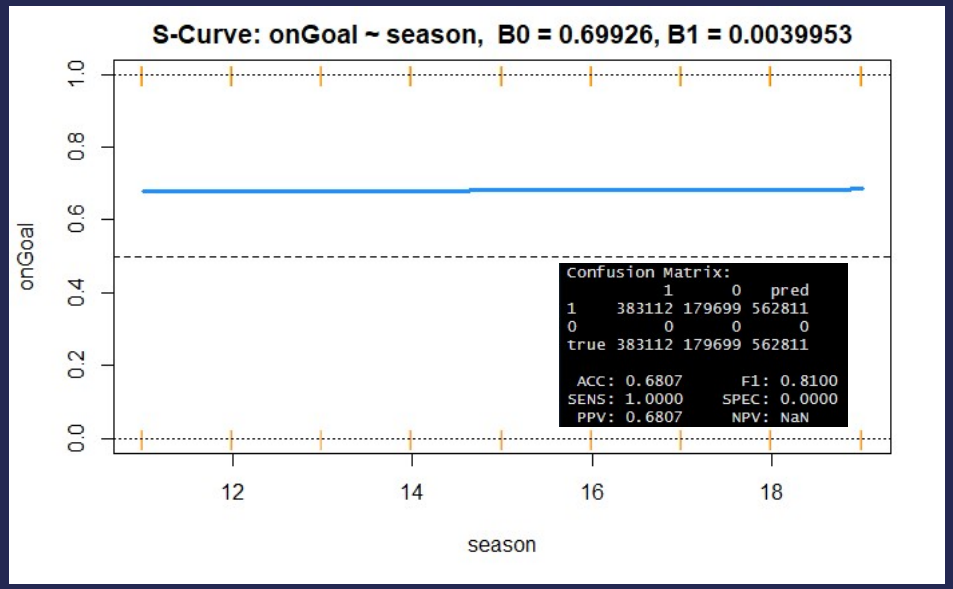
- **st_x** is extremely significant! S-curve midpoint is -18.146 (just outside the defensive zone)
- **season** is terrible; s-curve midpoint is -175.023 (calendar year 1825)



Simple Logistic Regression: season



- **season** is clearly not as strong as the original model suggests, at least not independently (sure is funny though)



Simple Logistic Regression: st_x

- Suggests NHL API arbitrarily defines blocked shots by distance; may be recategorized under "missed shots"