A simple logistic regression was used to create an "XG Model" for a hockey shot given these statistics: event type, period, time(in seconds), the distance of the shot, the angle with which the shot was taken, was the shot off of a rebound, was the shot off of a pass, is the shot a cross-line shot, is the shot a one-touch, and was the shot off of a fast attack. A logistic regression model was chosen because a logistic regression can be effective when the number of features is relatively small, and the dependent is discrete, or binary. Depending on the size of the data, a logistic regression could be an effective model to use because of its computational efficiency.

During my initial analysis of the data, there were a few things that stuck out to me as interesting data trends. I looked at the means of each feature with respect to goals scored and the results can be seen by the table below:

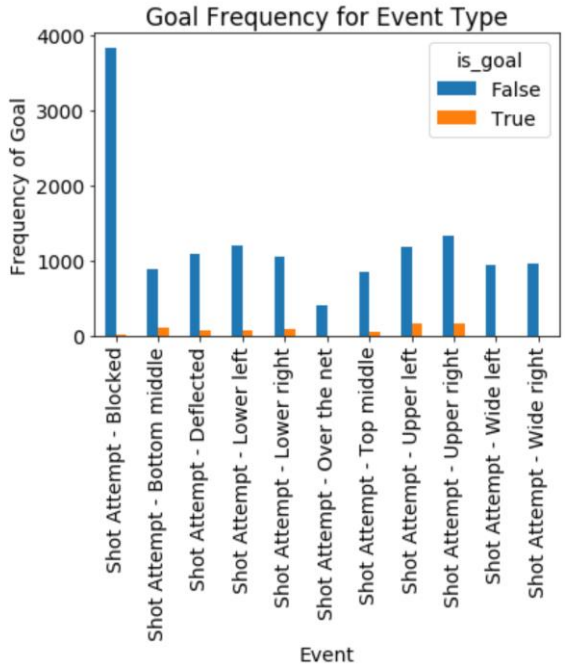
Shot Outcome	Period	Time(Seconds)	Distance	Angle	Is Rebound	Is Pass Before	Is Cross Line	Is One-Touch	Is Fast-Attack
Goal	2.0002	610.06	9.0	55.97	19.5%	23.65%	28.1%	46.2%	9.12%
No Goal	2.022	603.40	13.73	56.69	5.08%	16.93%	16.94%	25.22%	13.98%

What I find interesting about this data is it seems like a rebound, pass-before, cross-line, and one-touch impact the probability of a shot becoming a goal. Also, as expected the distance of the shot has an effect too. Another interesting observation is that the average angle of a goal and no goal are relatively similar. That doesn't mean that the angle of the shot has no effect on the probability of a shot becoming a goal, but it does call for more analysis in that statistic.

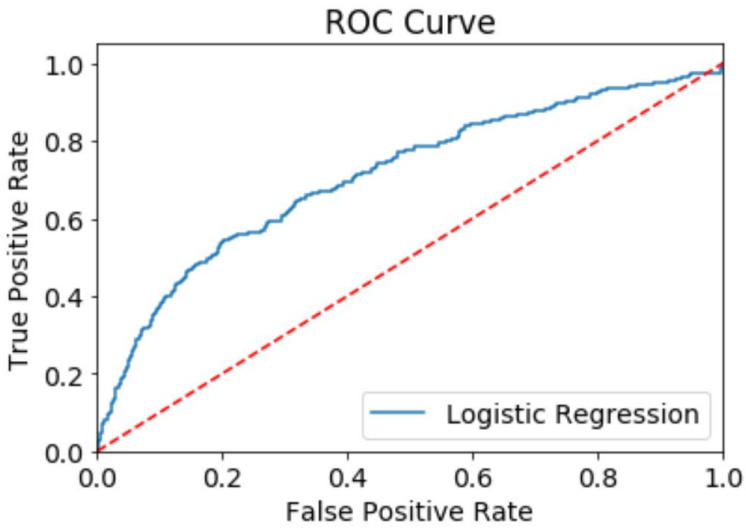
The next thing I did in my initial data analysis is run a logistic regression using the statsmodel library within python to see how different the weights would be between the scikit-learn library and statsmodel. Interestingly the only variable that was statistically significant (p-value below 0.05) was the rebound shot. Obviously this is incorrect because the common hockey fan would know that other shot types or events should be statistically significant as well so it calls for more analysis potentially using a different model.

```
         Current function value: 0.187252
         Iterations 8
                         Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.059
Dependent Variable: is_goal          AIC:              5414.6033
Date:               2019-01-21 19:16 BIC:              5482.7844
No. Observations:   14410            Log-Likelihood:   -2698.3
Df Model:           8                LL-Null:          -2868.0
Df Residuals:       14401            LLR p-value:      1.7222e-68
Converged:          1.0000           Scale:            1.0000
No. Iterations:     8.0000
----------------------------------------------------------------
                 Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
----------------------------------------------------------------
period           -0.3464   0.0377  -9.1833 0.0000 -0.4203 -0.2725
time             -0.0006   0.0001  -6.3318 0.0000 -0.0008 -0.0004
distance         -0.1697   0.0074 -22.8827 0.0000 -0.1842 -0.1552
angle            -0.0011   0.0013  -0.8023 0.4224 -0.0037  0.0015
is_rebound        0.3062   0.1466   2.0885 0.0368  0.0188  0.5936
is_pass_before    0.1682   0.1527   1.1013 0.2707 -0.1311  0.4674
is_cross_line     0.0925   0.0919   1.0066 0.3141 -0.0876  0.2725
is_one_touch      0.1057   0.1382   0.7644 0.4446 -0.1653  0.3766
is_fast_attack   -0.2405   0.1370  -1.7553 0.0792 -0.5089  0.0280
----------------------------------------------------------------
```

Another statistic I looked at was the goal frequency for event type. Because we don't have a categorical value assigned to the shot type it isn't possible to include it in our logistic regression model for expected goals. Out of the event types, it seemed like upper left and upper right shot types had the most significant impact on whether the shot became a goal or not.



Finally, I created a logistic regression model using the scikit-learn library within python. The model was trained off of 70% of the data and tested on the remaining 30%. The model performed with an overall accuracy of 95% which is obviously better than randomly predicting whether a shot will become a goal or not. The model correctly predicted 4,108 samples versus 215 incorrect predictions.



In conclusion, the logistic regression xg model seems to predict a shot's probability of becoming a goal rather effectively. However, next steps for this model would be to analyze the independent variables more either by reducing features or adding more that could be more statistically significant.