**NEW PROBABILITIES MODEL – DAN**

In the midterm project, we came up with a logit model to try and predict the probabilities of scoring a goal, given the (x, y) starting coordinates on the field, the (x, y) ending coordinates, as well as the distance travelled by the ball.

There were some issues with this approach:

1. There was high multicollinearity between the terms in the model. For example, “Distance” is dependent on the (x, y) coordinates for starting and ending plays.
2. The data from the independent variables was not centered.
3. The Betas obtained at the summary of the model were not the probabilities of scoring a goal; rather, they were the logit coefficients of the model. We had not seen the technique to convert the logit coefficients to actual probabilities (i.e., fitted values), so we were unable to propose a better model.
4. From the previous point, it is then necessary to build an additional vector in the database that includes reasonable values of the probabilities of scoring a goal. We then need to build an additional model that results in Betas expressed as probabilities of scoring a goal.

With this, we did 4 things:

1. Checked that the selection of observations is balanced: similar number of observations for success as for failure (1 or 0 in “Success” vector). These counts resulted in 30,475 observations of success vs 33,786 observations of failure. It’s relatively balanced, so no further action was taken.

1. We centered the values of independent variables: subtract the mean from of each variable from each observation of that variable.
2. We built new logit models, considering interaction terms.
3. We evaluated the logit models based on BIC and VIF criteria, and selected the one that had no multicollinearity and lowest BIC. Model 6 is expressed as

Note that the interaction expressed as is equivalent to

1. We included the probability vector to the database from the fitted values.

The summary statistics of the new vector is



1. We made a new model with the fitted values (probabilities of success) as our dependent variable.

The summary statistics of the new model (model 6b) are

