EDA - Assignment 7

Jivitesh Poojary

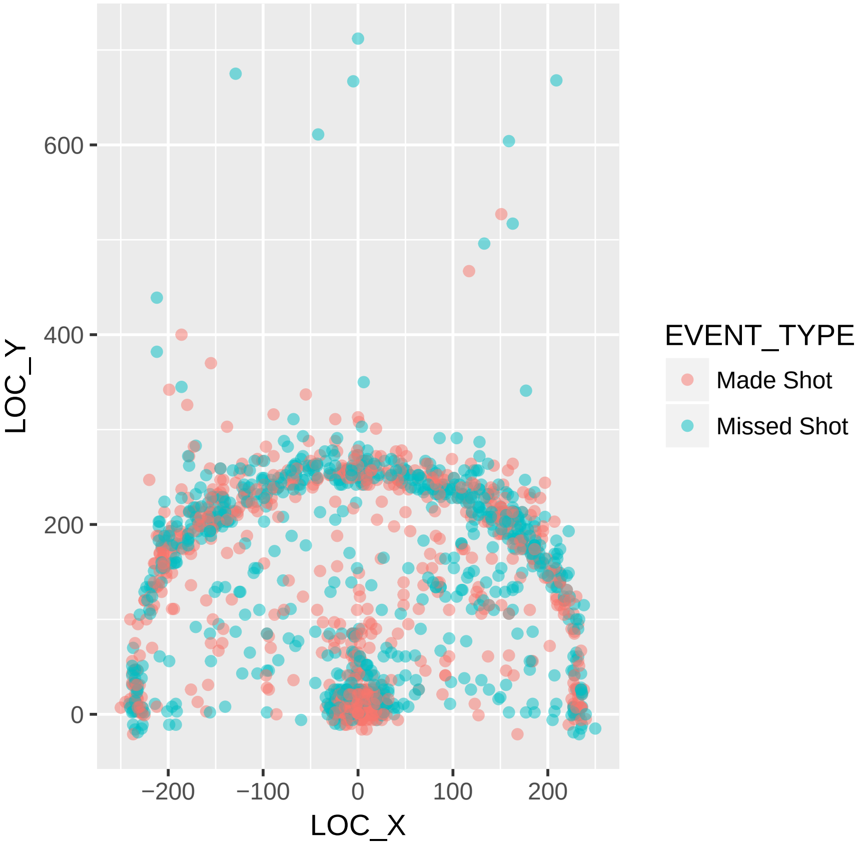
March 22, 2017

Answer 1

Q1. Plot the data to show the location of Curry's shots using color to distinguish between made and missed shots, similarly to the picture below. (You don't have to include the picture of the court unless you want to show off.) NB: It should use coord_fixed() since the units are the same for both axes.

ANSWER:

- Please find the plot below



Answer 2

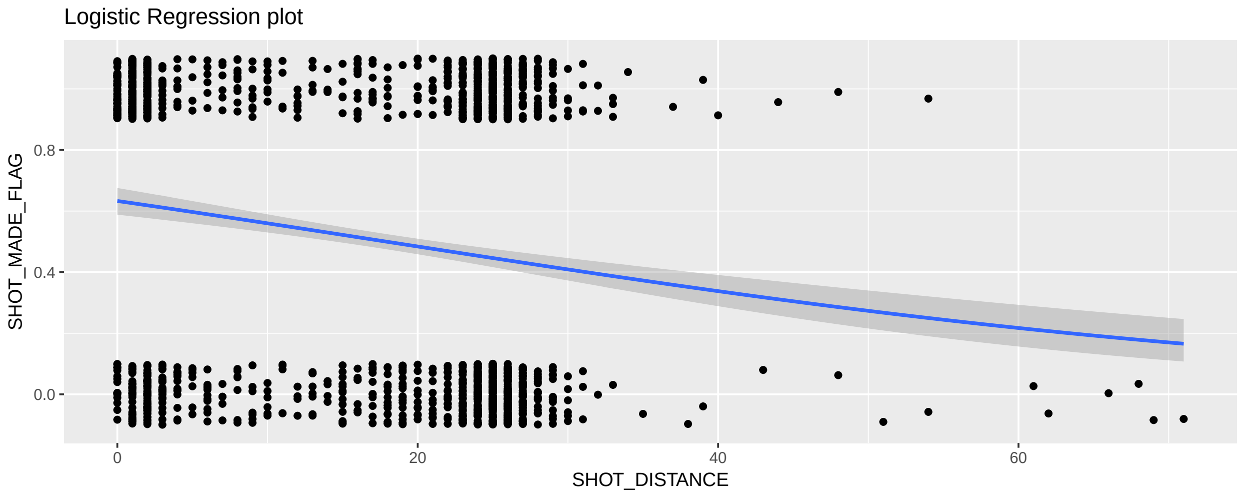
Q2. Fit a logistic regression to predict whether the shot is made, using the single predictor SHOT_DISTANCE. Draw an appropriate ggplot of the fitted curve and write an equation for the fit.

ANSWER:

- The Logistic Regression model is fitted using glm function in R.
- The plot below shows the blue line showing the line for model of logistic regression while the orange curve shows the fitted values for loess model
- The equation of the fitted value is : Logit(Probability of SHOT) = 0.54508 - 0.03045(SHOT_DISTANCE)

```
curry.logit = glm(SHOT_MADE_FLAG ~ SHOT_DISTANCE, family = binomial, data = curry)
summary(curry.logit)
```

```
##
## Call:
## glm(formula = SHOT_MADE_FLAG ~ SHOT_DISTANCE, family = binomial,
##     data = curry)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4159  -1.0996   0.9563   1.2309   1.6654
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.54508    0.09606   5.674 1.39e-08 ***
## SHOT_DISTANCE -0.03045    0.00467  -6.521 6.97e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2215.2  on 1597  degrees of freedom
## Residual deviance: 2171.0  on 1596  degrees of freedom
## AIC: 2175
##
## Number of Fisher Scoring iterations: 4
```

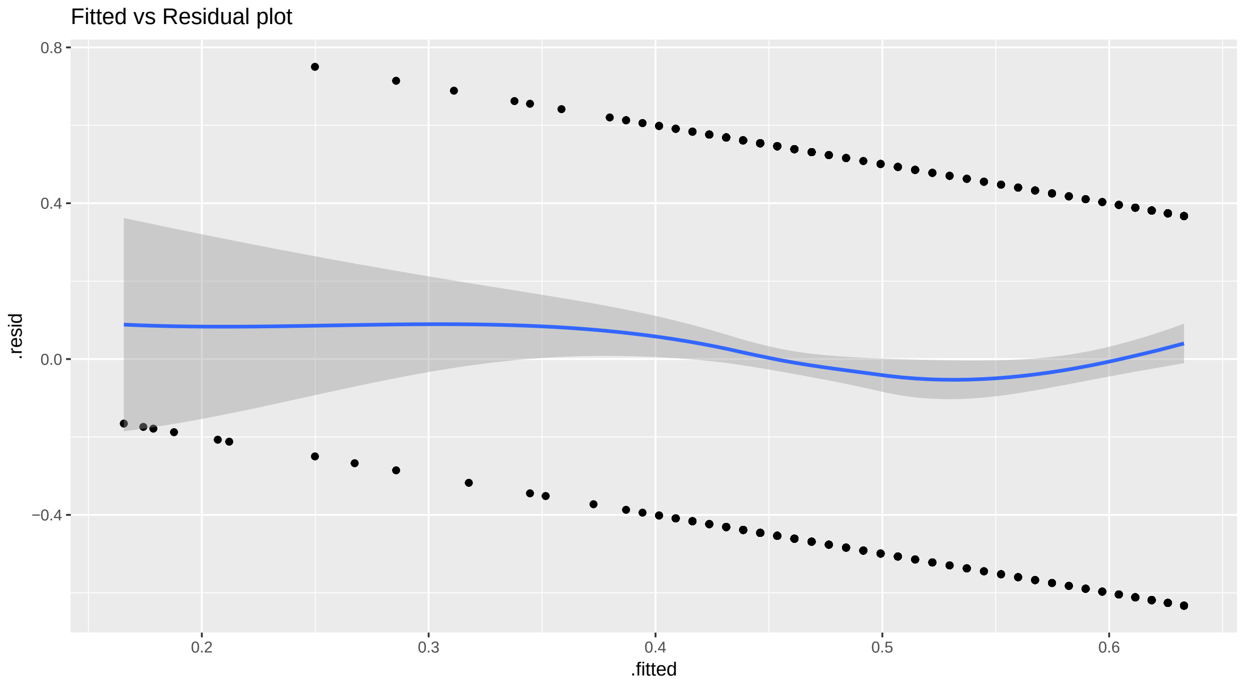


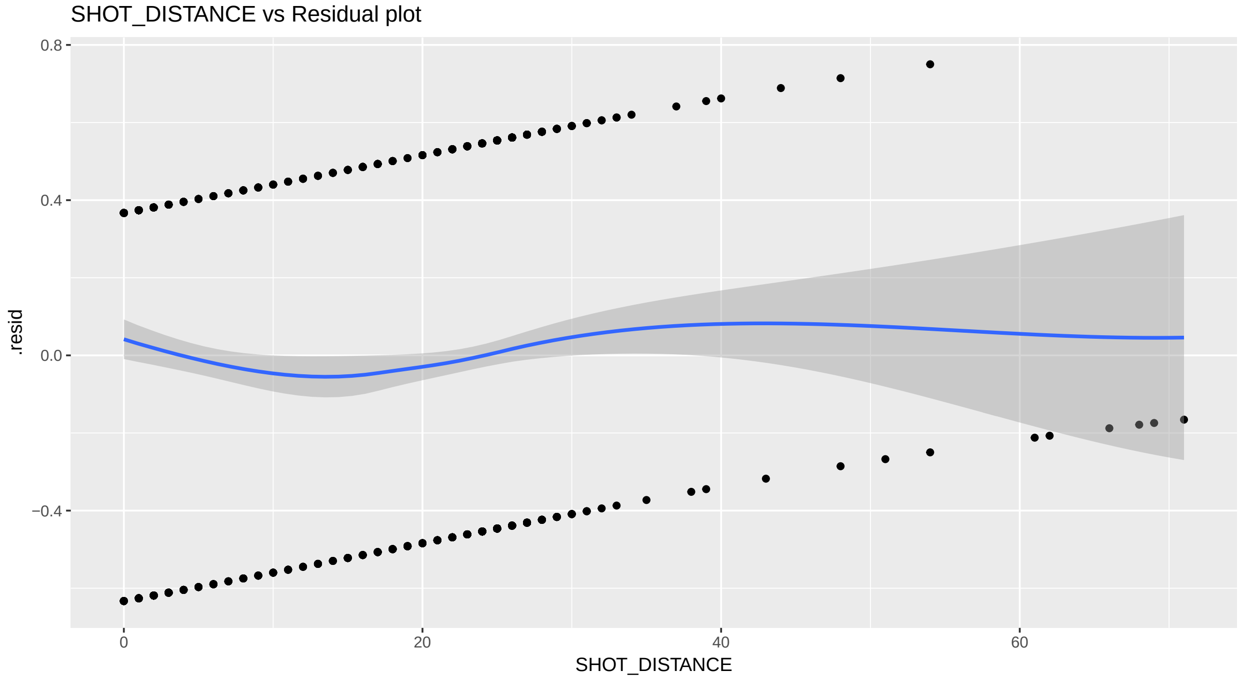
Answer 3

Q3. Plot the residuals in a way that shows where the logistic regression doesn't fit the data well. Describe in some detail how the model is inaccurate.

ANSWER:

- From the fitted vs residual plot below we can see that the model does not fit well.
- The SHOT_DISTANCE vs residual plot also shows the model is not a good fit
- The loess fit on the data is curve which has a large deviation around the residual '0' line





Answer 4

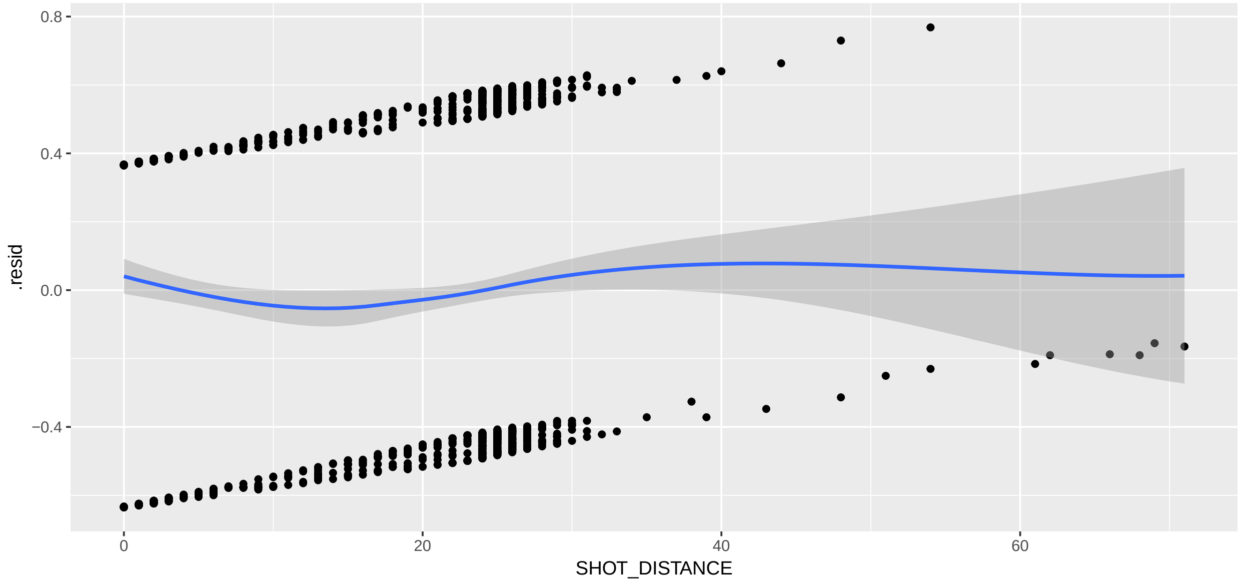
Q4. Fit a better model. You could try a different functional form or a model with more predictors (as long as you use the predictors sensibly.) Your model doesn't have to be perfect, just better. Draw a graph that shows how your model differs from the simple logistic regression, and convince us that your model is better.

ANSWER:

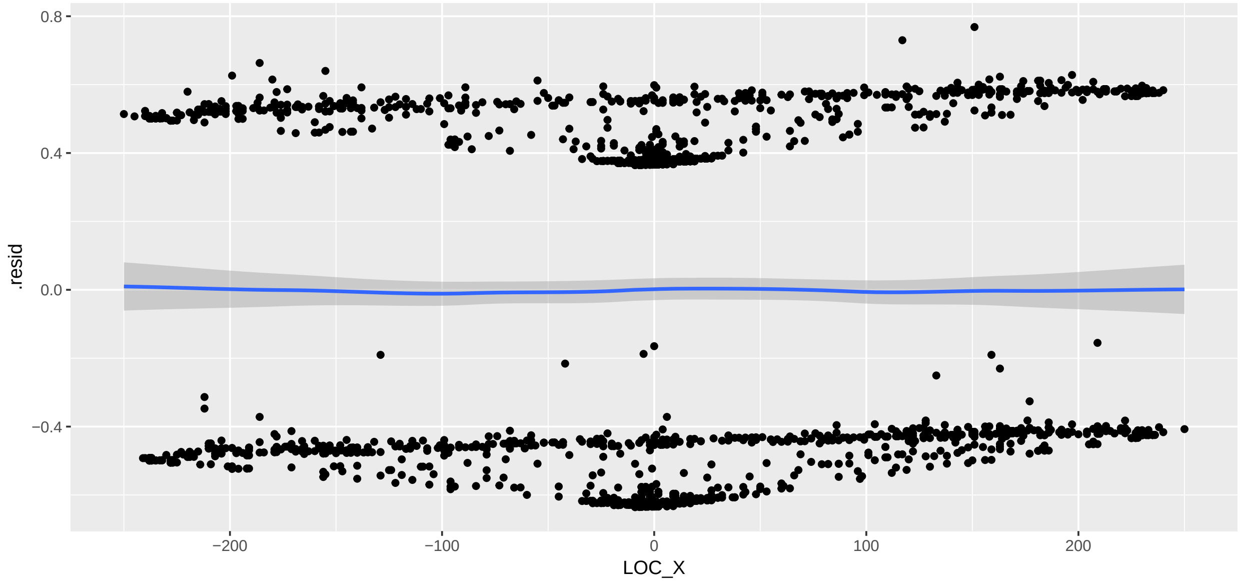
- This is a much better model because the Residual deviance residuces significantly compared to other models, the decrease is from 2171.0 to 2168.4. This is significant when we compare the Residual deviance of 2167.2 for the most complicated model having three way interaction. Thus along with an improved predictability the model is also simple to understand
- Another reason we may select this model is because only for this model the AIC value decreases compared to the most basic model. We have seen an increase in the AIC value for all other models.
- We have also seen the p-values for the variables to have an idea of their significance.
- Similarly, Another reason why this model is better than the first model is because the deviance values for this model are lower. Though this is not a great improvement but it is an improvement none the less.
- We have drawn some plots comparing the residuals with variables. There is no significant improvement in the residuals vs SHOT_ DISTANCE curve. It looks like the plot drawn earlier. The residuals vs LOC_X plot shows that the fitted line almost overlaps with the residual = '0' line which subtatiates the goodness of fit for the model.

```r
curry.logit2 = glm(SHOT_MADE_FLAG ~ SHOT_DISTANCE + LOC_X ,
                   family = binomial, data = curry)
summary(curry.logit2)
```

```
##
## Call:
## glm(formula = SHOT_MADE_FLAG ~ SHOT_DISTANCE + LOC_X, family = binomial,
##     data = curry)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.420  -1.116   0.953   1.199   1.710
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.5494219  0.0961619   5.714 1.11e-08 ***
## SHOT_DISTANCE -0.0305962  0.0046749  -6.545 5.96e-11 ***
## LOC_X         -0.0006382  0.0003964  -1.610    0.107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2215.2  on 1597  degrees of freedom
## Residual deviance: 2168.4  on 1595  degrees of freedom
## AIC: 2174.4
##
## Number of Fisher Scoring iterations: 4
```



SHOT_DISTANCE vs Residual plot



LOC_X vs Residual plot