DATA 621—Assignment no. 3

Critical Thinking Group 2

October 30, 2019

Contents

Executive Overview

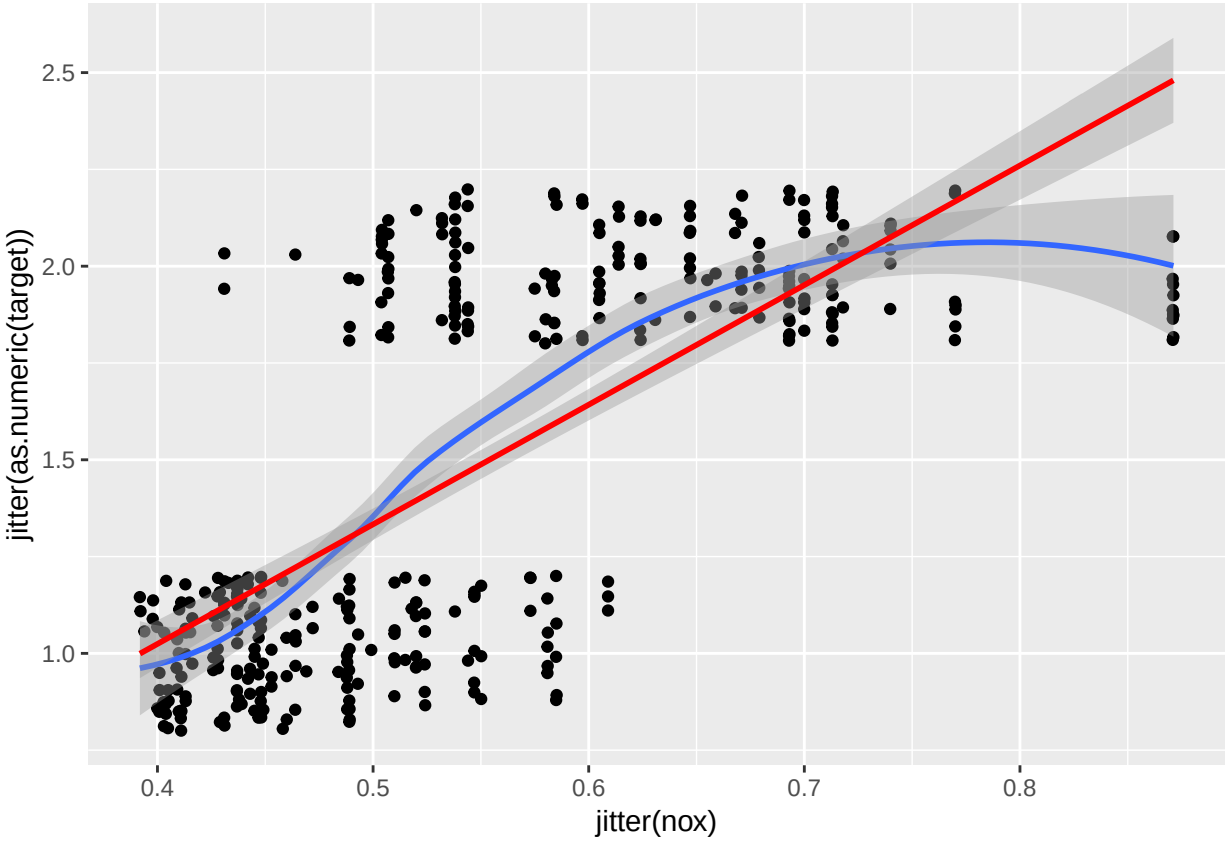
blah balh

Create train and test sets using the `caret` machine learning package:

Only use the `train` data frame until the very end of the process, when we use test to evaluate how effective the model is!

Data Exploration

```
ggplot(train, aes(x=jitter(nox), y=jitter(as.numeric(target)))) +
  geom_point() +
  geom_smooth() +
  geom_smooth(method='lm', color='red')
```



Data Preparation

Modeling

Baseline model, which just predicts the class proportion. It is the model to beat. If we are having trouble improving on this model, we know we are doing something wrong.

This dummy model has an accuracy of about 0.50, sensitivity of 1, and specificity of 0. Since it has zero predictive power, we know that it has a pseudo-R^2 of 0.

```
m_0 <- glm(target ~ 1, train, family=binomial())
pred_0 <- factor(round(predict(m_0, train, type='response')), levels=c('0', '1'))
confusionMatrix(data=pred_0, reference=train$target)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 190 184
##          1   0   0
##
##                Accuracy : 0.508
##                  95% CI : (0.4561, 0.5598)
##     No Information Rate : 0.508
##     P-Value [Acc > NIR] : 0.5207
```

```
##
##                  Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.000
##             Specificity : 0.000
##          Pos Pred Value : 0.508
##          Neg Pred Value :   NaN
##              Prevalence : 0.508
##          Detection Rate : 0.508
##    Detection Prevalence : 1.000
##       Balanced Accuracy : 0.500
##
##        'Positive' Class : 0
##
```