● Use the Naive Bayes model to predict whether the flight is delayed or not. Use only categorical variables for the predictor variables. Note that Week and Time variables need to recoded as factors.

Recode the Departure time:

data.CRS\_DEP\_TIME = data.CRS\_DEP\_TIME / 100

data['CRS\_DEP\_TIME'] = data['CRS\_DEP\_TIME'].transform(lambda i: math.floor(i))

array([14, 16, 12, 17, 10, 8, 21, 9, 20, 15, 6, 18, 13, 19, 11, 7])

Length: 16

Recode categorical data:

import category\_encoders as ce

encoder = ce.OneHotEncoder(cols=['CARRIER', 'DEST', 'ORIGIN', 'DAY\_WEEK'], use\_cat\_names=True)

data = encoder.fit\_transform(data)

CRS\_DEP\_TIME CARRIER\_OH CARRIER\_DH CARRIER\_DL CARRIER\_MQ CARRIER\_UA CARRIER\_US CARRIER\_RU CARRIER\_CO DEST\_JFK DEST\_LGA DEST\_EWR ORIGIN\_BWI ORIGIN\_DCA ORIGIN\_IAD DAY\_WEEK\_4.0 DAY\_WEEK\_5.0 DAY\_WEEK\_6.0 DAY\_WEEK\_7.0 DAY\_WEEK\_1.0 DAY\_WEEK\_2.0 DAY\_WEEK\_3.0 Flight Status

● Output both a counts table and a proportion table outlining how many and what proportion of flights were delayed and on-time at each of the three airports.

Total flights BWI: 145

Total flights DCA: 1370

Total flights IAD: 686

-------------------------------------------------

Delayed flights BWI: 37

Delayed flights DCA: 221

Delayed flights IAD: 170

-------------------------------------------------

Proportion Delayed flights BWI: 0.25517241379310346

Proportion Delayed flights DCA: 0.16131386861313868

Proportion Delayed flights IAD: 0.2478134110787172

● Output the confusion matrix and ROC for the validation data

Confusion matrix, without normalization

[[504 208]

[ 88 81]]

Normalized confusion matrix

[[0.71 0.29]

[0.52 0.48]]





