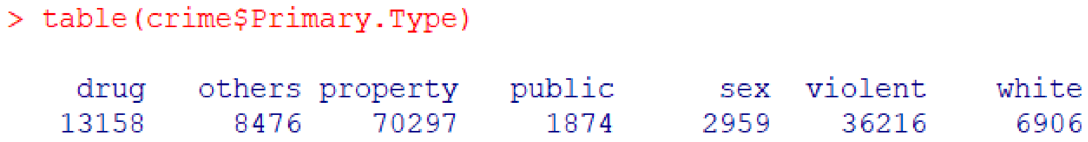
CHICAGO CRIME DATA

KUBIG
박소현 김효익 조송현 조규선 이영신

Imbalanced Data

Problems with Imbalanced Data

Category	Type of Crime
Violent	"ASSAULT”, "BATTERY”, "HOMICIDE”, "INTIMIDATION”, "KIDNAPPING”, "CONCEALED CARRY LICENSE VIOLATION”, "WEAPONS VIOLATION”
Property	"ARSON”, "BURGLARY”, "CRIMINAL DAMAGE”, "CRIMINAL TRESPASS”,"MOTOR VEHICLE THEFT”, "ROBBERY”, "THEFT”
Sex	"CRIM SEXUAL ASSAULT”, "OFFENSE INVOLVING CHILDREN”, "PROSTITUTION”, "SEX OFFENSE”, "STALKING”
White	"DECEPTIVE PRACTICE”, "GAMBLING”
Public	"INTERFERENCE WITH PUBLIC OFFICER”,"OBSCENITY”, "PUBLIC INDECENCY”,"PUBLIC PEACE VIOLATION”
Drug	"LIQUOR LAW VIOLATION”, "NARCOTICS”, "OTHER NARCOTIC VIOLATION”
Others	"NON - CRIMINAL”, " NON - CRIMINAL”, "OTHER OFFENSE”

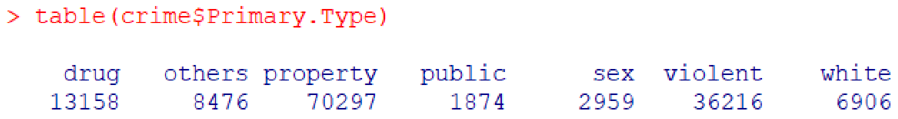


```
> table(crime$Primary.Type)

    drug   others property   public      sex  violent    white 
   13158     8476    70297     1874     2959    36216     6906 
```

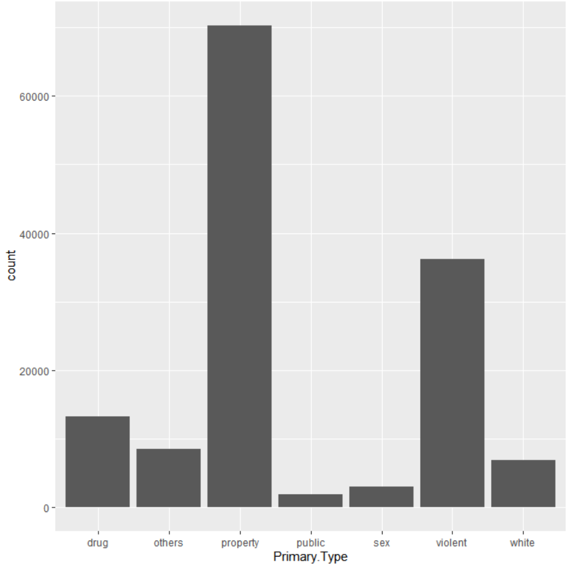
Problems with Imbalanced Data

Category	Type of Crime
Violent	"ASSAULT", "BATTERY", "HOMICIDE", "INTIMIDATION", "KIDNAPPING", "CONCEALED CARRY LICENSE VIOLATION", "WEAPONS VIOLATION"
Property	"ARSON", "BURGLARY", "CRIMINAL DAMAGE", "CRIMINAL TRESPASS","MOTOR VEHICLE THEFT", "ROBBERY", "THEFT"
Sex	"CRIM SEXUAL ASSAULT", "OFFENSE INVOLVING CHILDREN", "PROSTITUTION", "SEX OFFENSE", "STALKING"
White	"DECEPTIVE PRACTICE", "GAMBLING"
Public	"INTERFERENCE WITH PUBLIC OFFICER","OBSCENITY", "PUBLIC INDECENCY","PUBLIC PEACE VIOLATION"
Drug	"LIQUOR LAW VIOLATION", "NARCOTICS", "OTHER NARCOTIC VIOLATION"
Others	"NON - CRIMINAL", " NON - CRIMINAL", "OTHER OFFENSE"

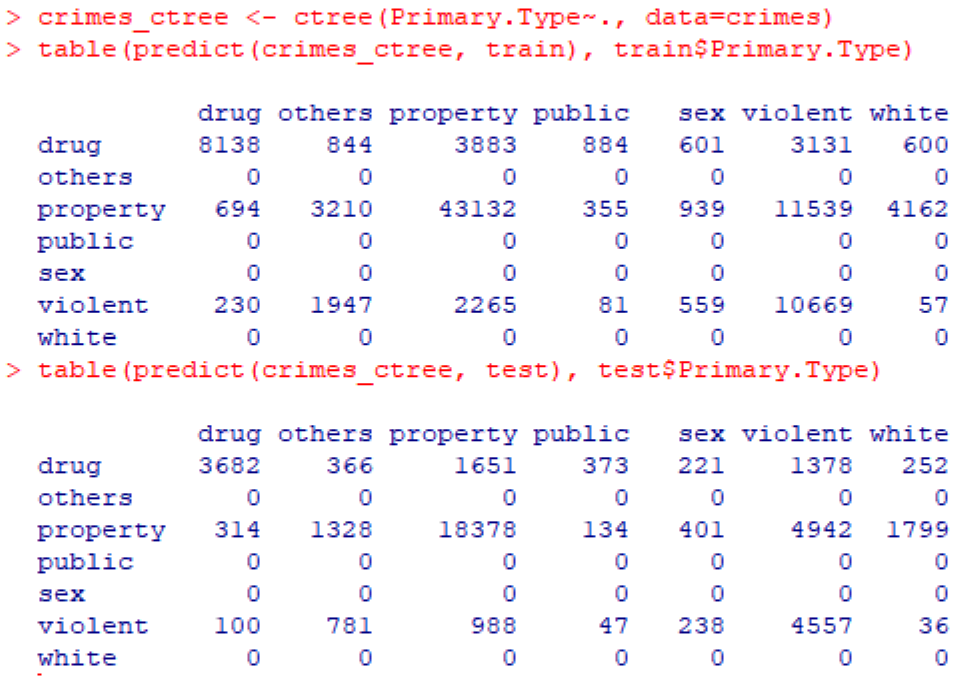


```
> table(crime$Primary.Type)

    drug   others property   public      sex  violent    white
   13158     8476    70297     1874     2959    36216     6906
```



Problems with Imbalanced Data



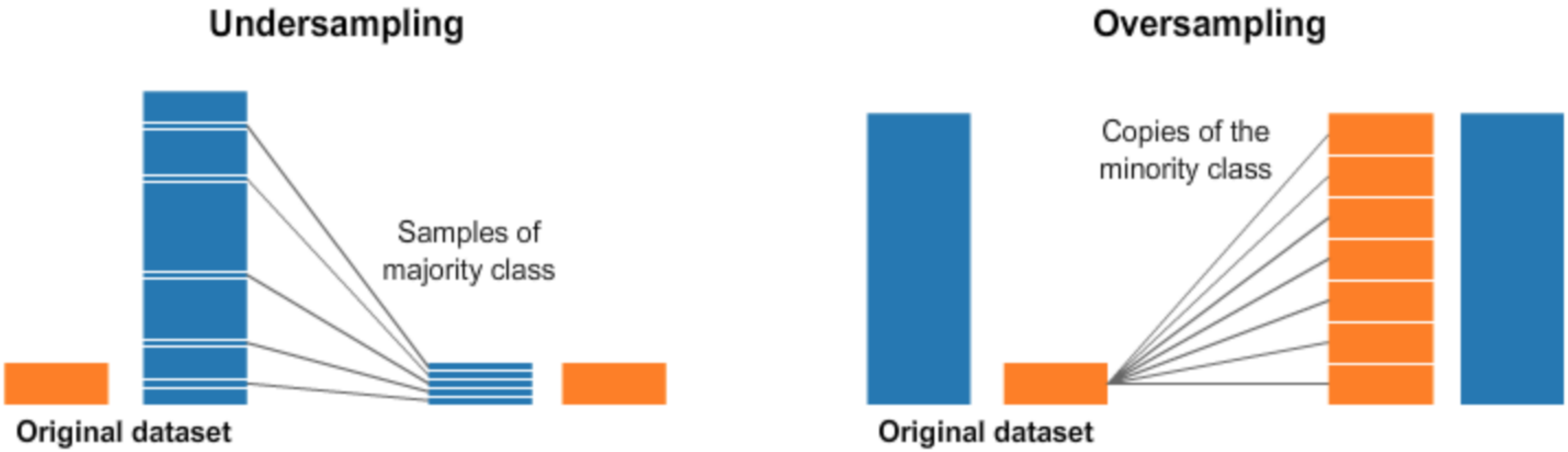
```
> crimes_ctree <- ctree(Primary.Type~., data=crimes)
> table(predict(crimes_ctree, train), train$Primary.Type)

           drug others property public   sex violent white
  drug     8138    844     3883    884   601    3131   600
  others      0      0        0      0     0       0     0
  property  694   3210    43132    355   939   11539  4162
  public      0      0        0      0     0       0     0
  sex         0      0        0      0     0       0     0
  violent   230   1947     2265     81   559   10669    57
  white       0      0        0      0     0       0     0
> table(predict(crimes_ctree, test), test$Primary.Type)

           drug others property public   sex violent white
  drug     3682    366     1651    373   221    1378   252
  others      0      0        0      0     0       0     0
  property  314   1328    18378    134   401    4942  1799
  public      0      0        0      0     0       0     0
  sex         0      0        0      0     0       0     0
  violent   100    781      988     47   238    4557    36
  white       0      0        0      0     0       0     0
```

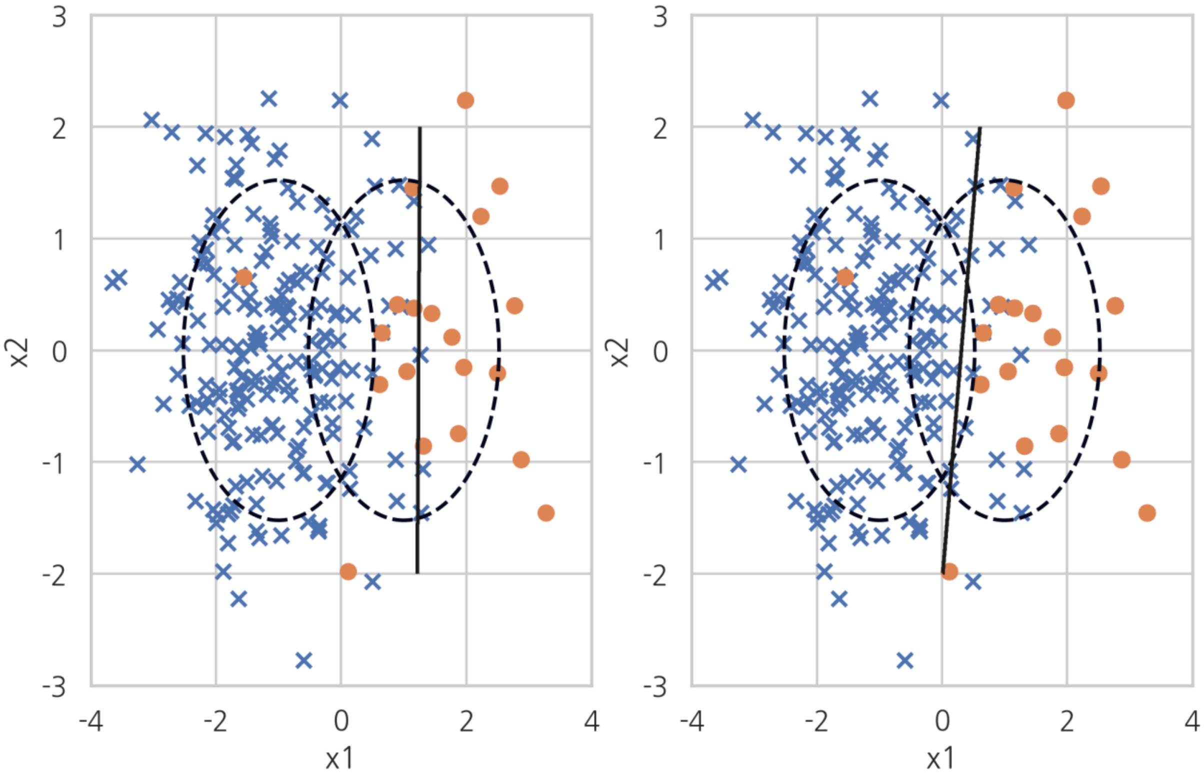
Oversampling vs Undersampling

Oversampling vs Undersampling

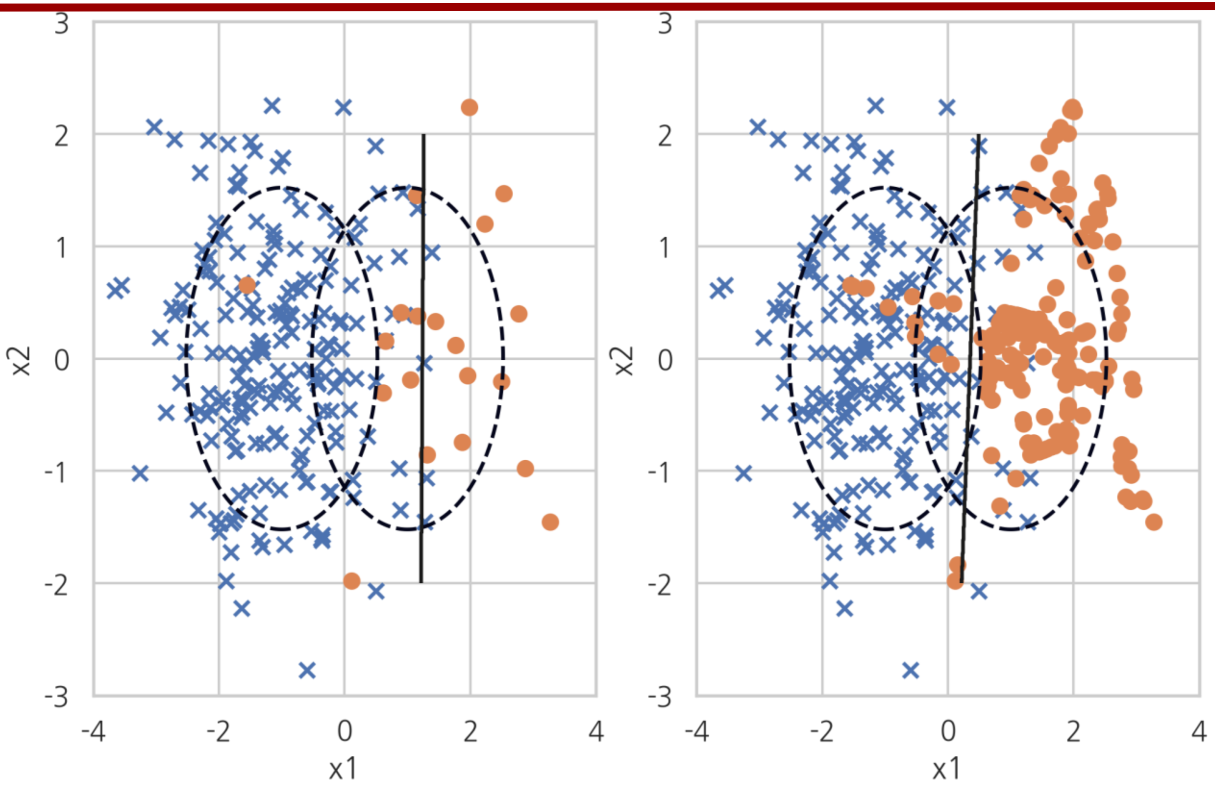


Oversampling

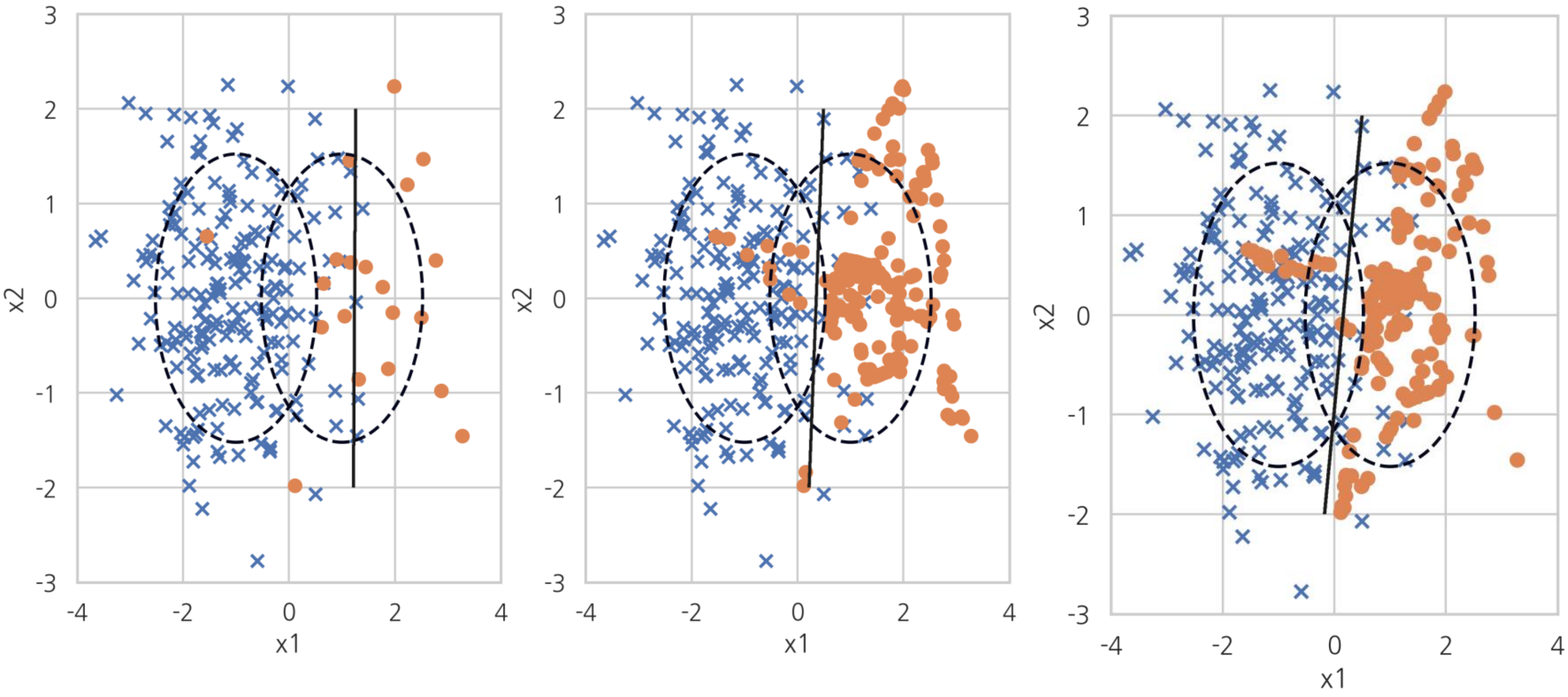
Random Oversampling



SMOTE (Synthetic Minority Oversampling Technique)

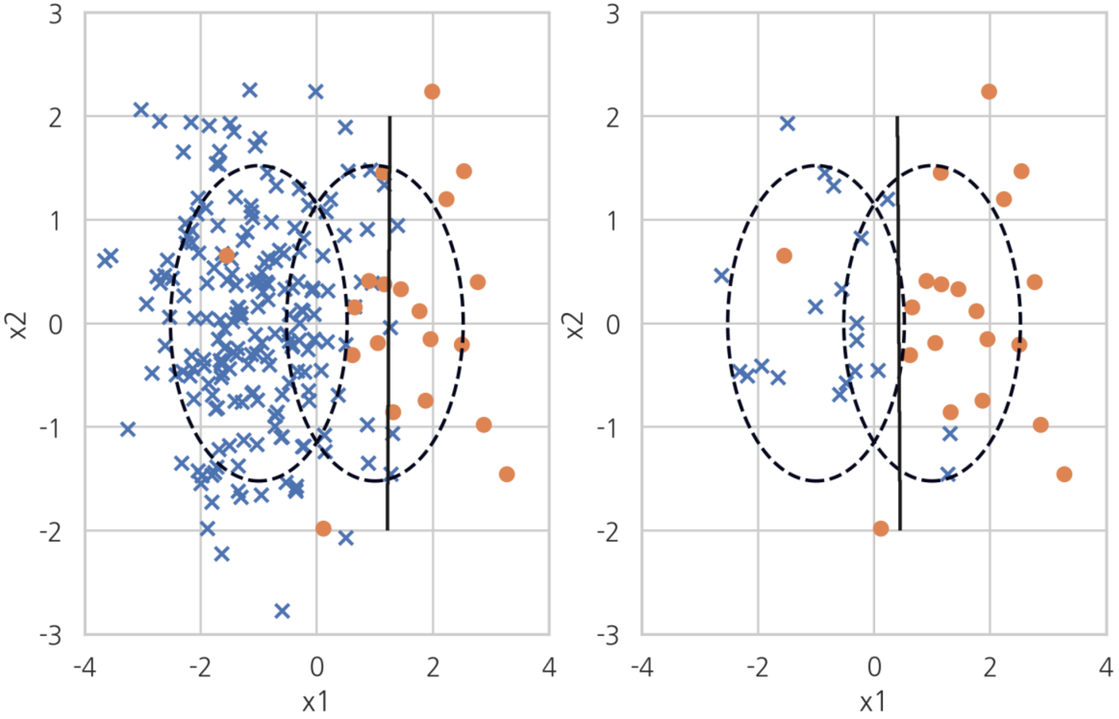


ADASYN

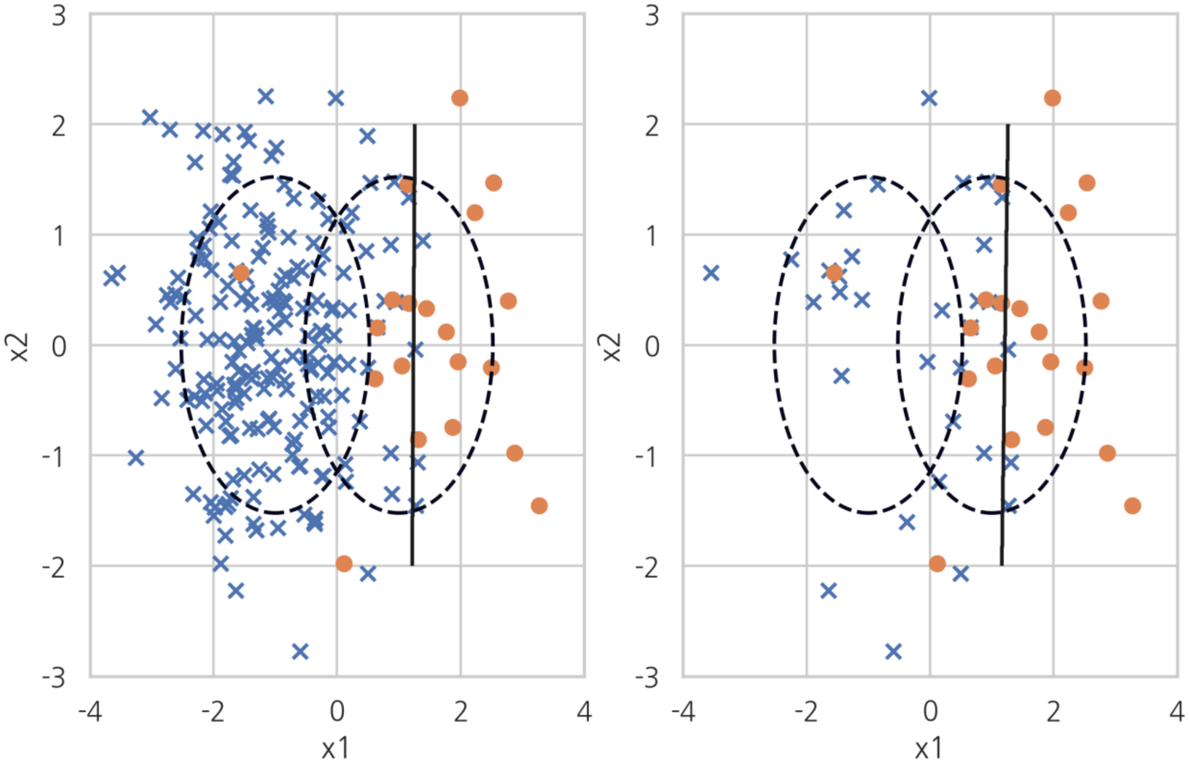


Undersampling

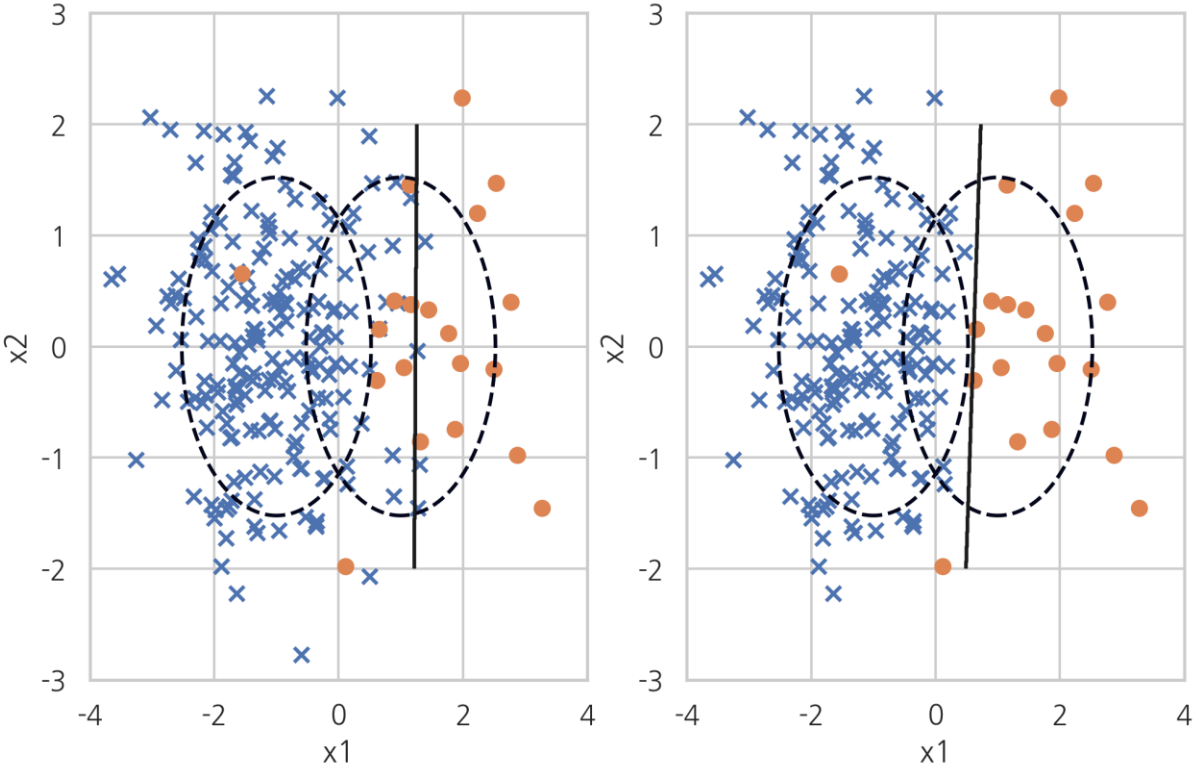
Random Under-Sampling



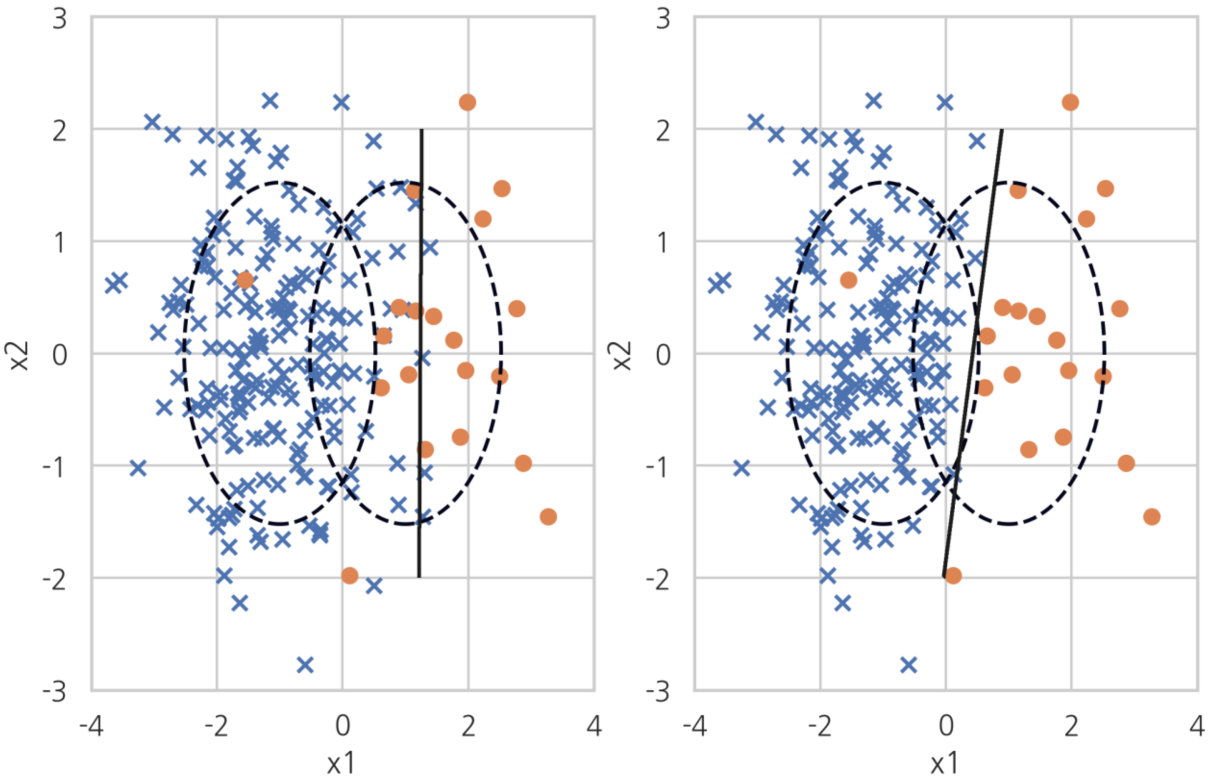
CNN(Condensed Nearest Neighbor)



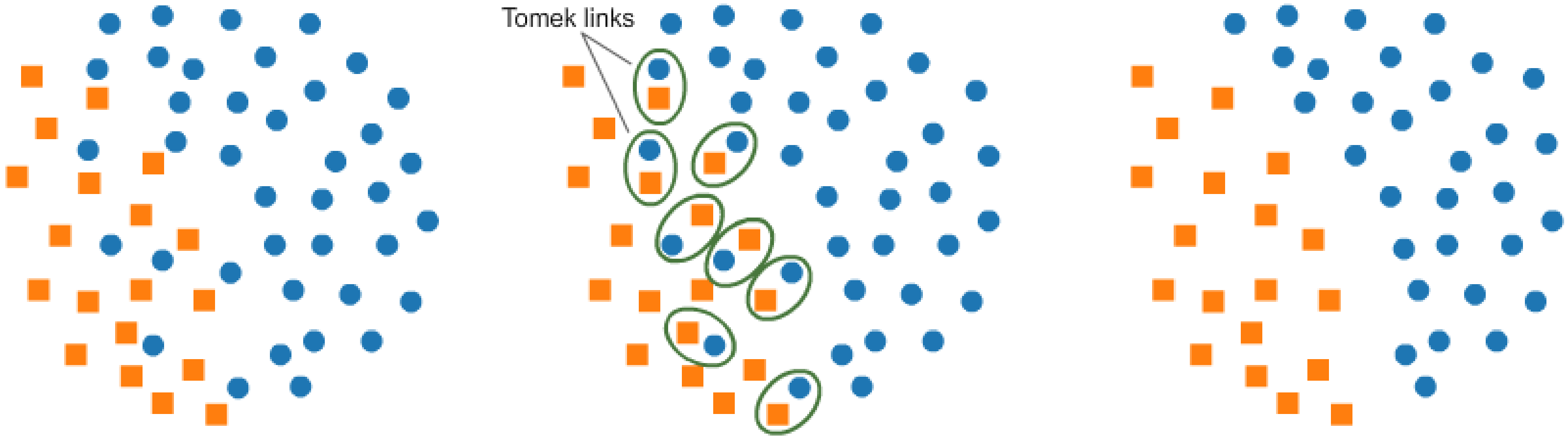
ENN(Edited Nearest Neighbor)



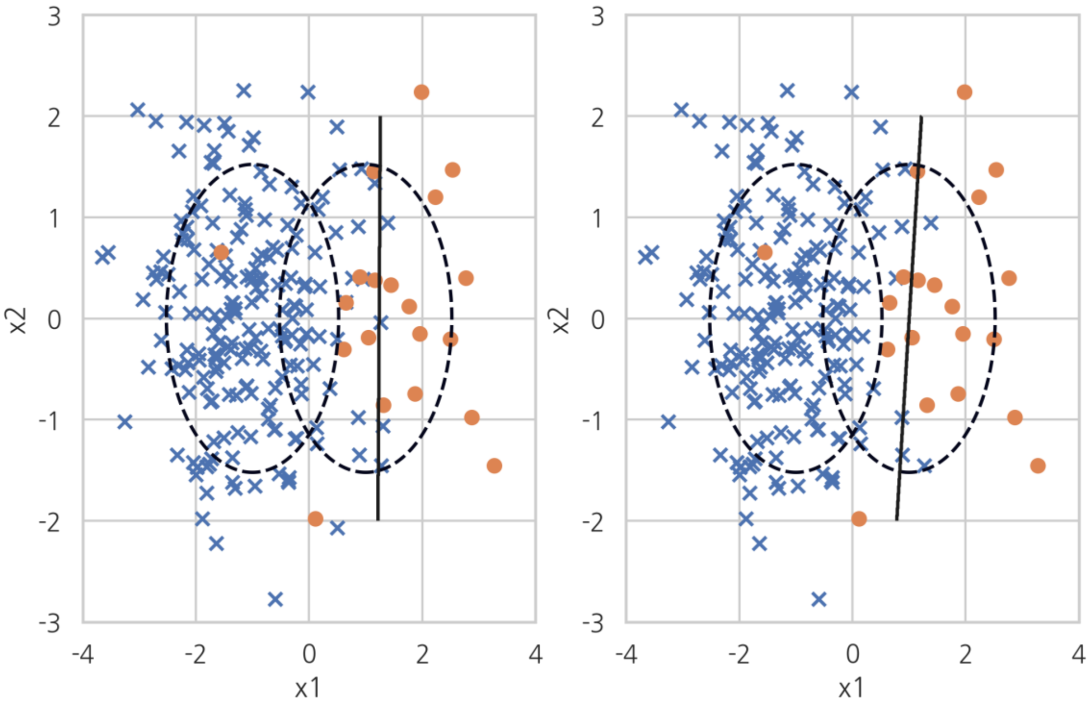
NCL (Neighborhood Cleansing Rule)



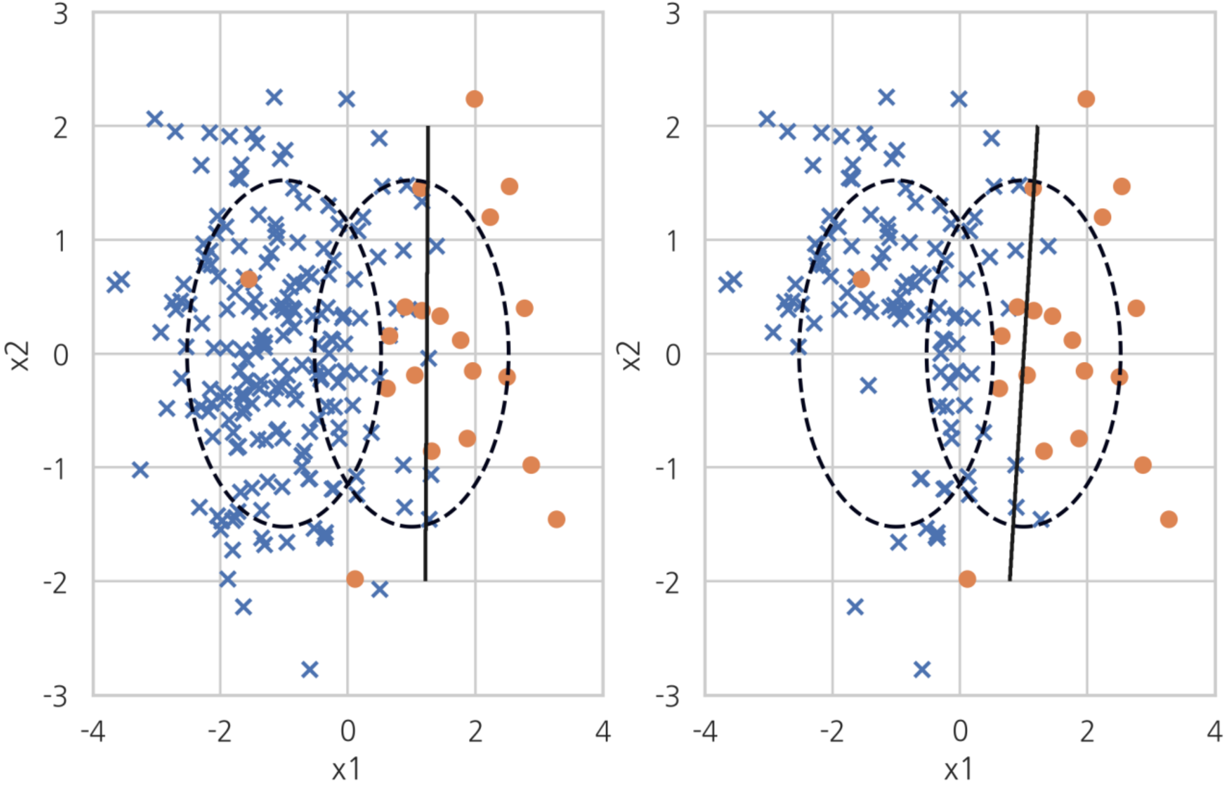
Tomek link Method



Tomek link Method



One Sided Selection



THANK YOU