NEGATIVE IMPACT:

BUSINESS DEPARTMENT:

SCENARIO 1:

If someone previously insured his/her car, %34 percent of them are not going to insure it again (but maybe in our company)

SCENARIO 2:

a)

first year == customer, [low price],

second year == customer, [high price] [reject, response = 0]

b)

first year == customer, [low price], [accident],

second year == customer, [high price] [reject, response = 0]

SCENARIO 3:

our marketing department is focusing on the customer who has not previously insured

Vehicle\_Damage 0 1

Previously\_Insured 0 1 0 1

Response

0 23080 164634 136849 9836

1 910 72 45642 86

# sample = summary() = ozet (kisaltilmis)

# 2k-5k (for visualization)

**NOTE: Sampling may cause problems if the “n” is small, OR, “data is highly imbalanced”**

balance = denge

imbalanced = dengesiz (cok fazla, cok az)

1-Annual premium’da, sabit bir fiyat var! (minimum fee)

2-30k ortalamasi olan bir normal dagilim

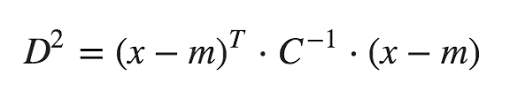
3-70 bin ustunde olan bir kisim var (outlier)

mahalanobis distance formula

2 different formulas:

1- euclidean / standard deviation (standardization)

2 – euclidean / correlation matrix



sqrt( (x-m)2 ) == euclidean

C= correlation matrix

-1

INVERSE

Age → azaldikta, Height → azaliyorsa [ positive strong correlation ] [+ yuksek magnitude ]

Ahmet => age:10, height: 120, iq = 110

Mehmet => age:5, height: 75, iq = 200

consider, there is no correlation between age and iq

Ahmet – Mehmet ===> ((120-75) + (10-5)) / CORRELATION (age, height)

distance normalization

because they are correlated, we do not need to sum up exactly two distances!

Ahmet – Mehmet (age, iq) ==> (10-5) + (200-110)

length(cm), length(inch)

object1 2.5 1

object2 5 2

object1 – object2 = (5-2.5) + (2-1)

AFTER NOON

data filling,

data ==> nulls , fill !!

1: Take our dataset, remove some values (cells) randomly (%10)

2: Try filling the empties/nulls

3: Measure how much same or similar value that original