Exploring anonymized data

Video overview

1. What is anonymized data?
2. What can we do with it?

Anonymized data

Anonymized data

Text	Encoded text
I want this table	7ugy 972h 98ww hj34
Table is what I want	hj34 4f08 rtte 7ugy 972h
This table is red	98ww hj34 4f08 4rj9
And this is me	jk8r 98ww 4f08 9jo4

Anonymized data

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmsp6u	1	NO	12.0	12	m268i97y
4	13e5dpzp	0	NO	140.12	14	m268i97y

Anonymized data

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmsp6u	1	NO	12.0	12	m268i97y
4	13e5dpzp	0	NO	140.12	14	m268i97y

- Explore individual features
  - Guess the meaning of the columns
  - Guess the types of the column
- Explore feature relations
  - Find relations between pairs
  - Find feature groups

Anonymized data

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmsp6u	1	NO	12.0	12	m268i97y
4	13e5dpzp	0	NO	140.12	14	m268i97y

- Explore individual features
  - Guess the meaning of the columns
  - Guess the types of the column
- Explore feature relations
  - Find relations between pairs
  - Find feature groups

Notebook

Notebook for video3에 상세내용 있음

random forest로 중요한 feature 선택

train.x8.mean(), train.x8.std(), train.x8.value_counts()

take a look at difference between neighbouring values in x8 - - > The most of the diffs are 0.04332159! - - > 원래 1이라고 가정 - - > 모든 값을 0.04332159으로 나눠줌

Ok, now we see .102468 in every value
this looks like a part of a mean that was subtracted during standard scaling
If we subtract it, the values become almost integers

let's round them - >Ok, what's next? In fact it is not obvious how to find shift parameter,

x8_int.value_counts()
do you see this - 1968? Doesn't it look like a year? ... So my hypothesis is that this feature is a year of birth!

위의 경우는 일반적이지 않음. 보통은 최소한 feature의 dtype 확인 - - > df.types(), df.info(), x.value_counts()

Exploring individual features: guessing types

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmsp6u	1	NO	12.0	12	m268i97y
4	13e5dpzp	0	NO	140.12	14	m268i97y

Exploring individual features: guessing types

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmsp6u	1	NO	12.0	12	m268i97y
4	13e5dpzp	0	NO	140.12	14	m268i97y

Helpful functions:

```
df.dtypes
df.info()
x.value_counts()
x.isnull()
```

Conclusion

- Two things to do with anonymized features:
  - **Try to decode the features**
    - Guess the true meaning of the feature
  - **Guess the feature types**
    - Each type needs its own preprocessing