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

First, we can try to decode or de-anonymize the data, in a legal way of course. That is, we can try to guess true meaning of the features. Sometimes de-anonymization is not possible, but what we almost surely can do, is to guess the type of the features,

- 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

EDA_video3.ipynb
How to decode the feature

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