PREPROCESSING FOR MACHINE LEARNING IN PYTHON

UFOs and prepocessing

Sarah Guido
Senior Data Scientist

Identifying areas for preprocessing

Important concepts to remember

- Missing data: dropna() and notnull()
- Types: astype()
- Stratified sampling: train_test_split(X, y, stratify=y)

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Categorical variables and standardization

Sarah Guido
Senior Data Scientist

Categorical variables

```
    state country       type
295    az      us      light
296    tx      us  formation
297    nv      us   fireball
```

- One-hot encoding: `pd.get_dummies()`

Standardization

- var()
- np.log()

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Engineering new features

Sarah Guido
Senior Data Scientist

UFO feature engineering

date	length_of_time	desc
6/16/2013 21:00	5 minutes	Sabino Canyon Tucson Arizona night UFO sighting.
9/12/2005 22:35	5 minutes	Star like objects hovering in sky&#44 slowly m...
12/31/2013 22:25	3 minutes	Three orange fireballs spotted by witness in E...

- Dates: .month or .hour attributes
- Regex: \d and .group()
- Text: tf-idf and TfidfVectorizer

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Feature selection and modeling

Sarah Guido
Senior Data Scientist

Feature selection and modeling

- Redundant features
- Text vector

Final thoughts

- Iterative processes
- Know your dataset
- Understand your modeling task

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Congratulations!

Sarah Guido
Senior Data Scientist