

UFOs and preprocessing

PREPROCESSING FOR MACHINE LEARNING IN PYTHON



James Chapman
Curriculum Manager, DataCamp

Identifying areas for preprocessing



Important concepts to remember

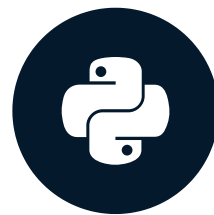
- Missing data: `.dropna()` and `.isna()`
- Types: `.astype()`
- Stratified sampling: `train_test_split(X, y, stratify=y)`

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Categorical variables and standardization

PREPROCESSING FOR MACHINE LEARNING IN PYTHON



James Chapman
Curriculum Manager, DataCamp

Categorical variables

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

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

Standardization

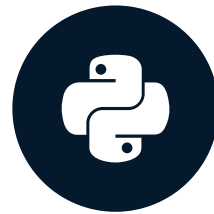
- `.var()`
- `np.log()`

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Engineering new features

PREPROCESSING FOR MACHINE LEARNING IN PYTHON



James Chapman
Curriculum Manager, DataCamp

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, slowly m...
12/31/2013 22:25	3 minutes	Three orange fireballs spotted by witness in E...

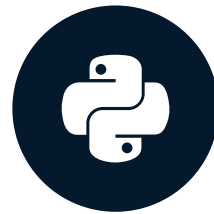
- Dates: `.dt.month` or `.dt.hour` attributes
- Regex: `\d` and `.group()`
- Text: tf-idf and `TfidfVectorizer`

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Feature selection and modeling

PREPROCESSING FOR MACHINE LEARNING IN PYTHON



James Chapman
Curriculum Manager, DataCamp

Feature selection and modeling

- Redundant features
- Text vector

Final thoughts

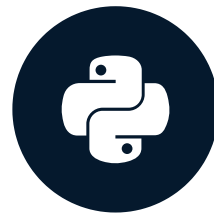
- Iterative processes
- Know your dataset
- Understand your modeling task

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Congratulations!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON



James Chapman

Curriculum Manager, DataCamp

What you've learned

- Preparing data for modeling:
 - Missing data
 - Incorrect types
 - Standardize numerical values
 - Process categorical values
 - Feature engineering
 - Select features for modeling

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON