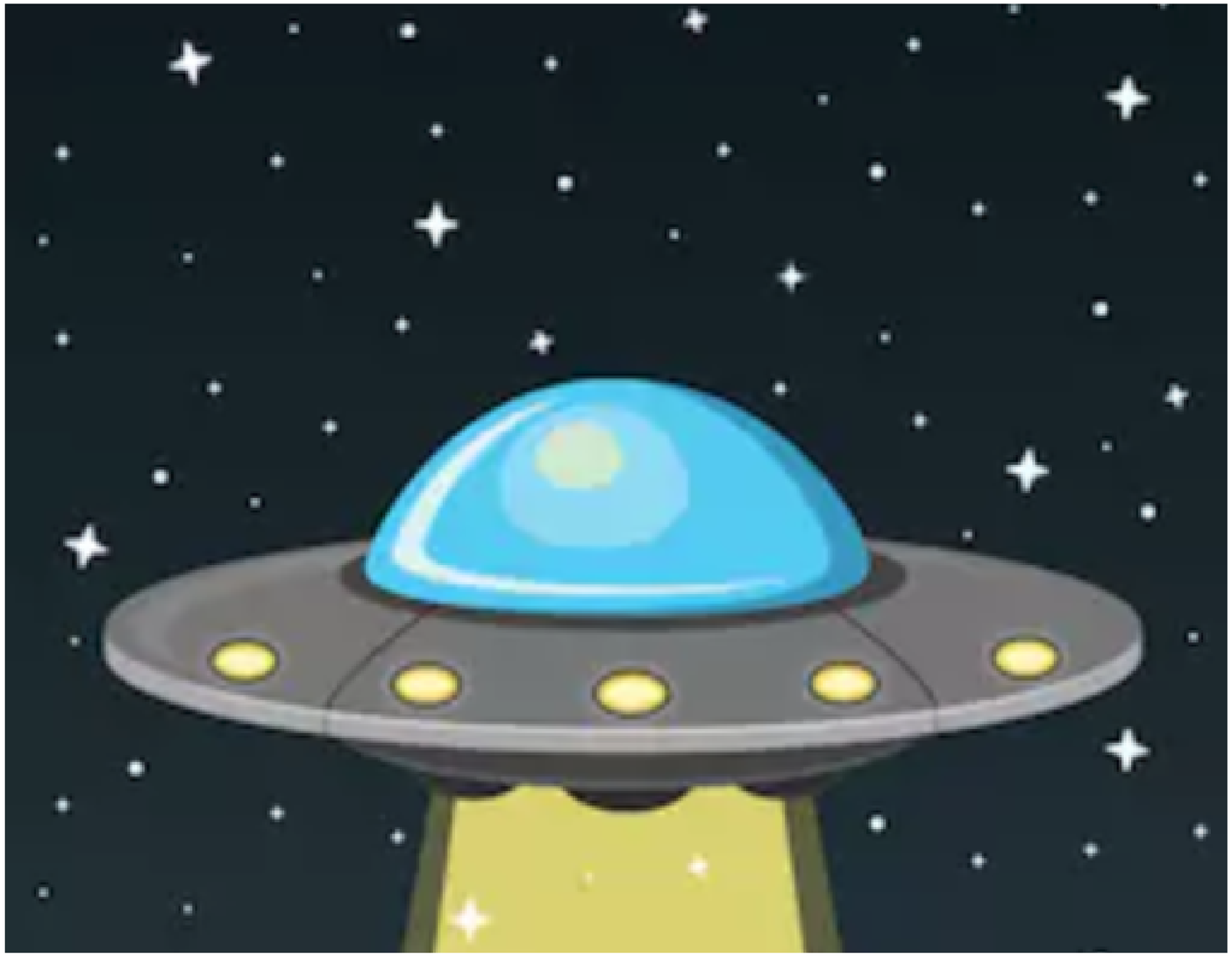
UFOs and preprocessing

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Sarah Guido
Senior Data Scientist

datacamp

Identifying areas for preprocessing



Important concepts to remember

- Missing data: `dropna()` and `notnull()`
- 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

Sarah Guido
Senior Data Scientist

datacamp

Categorical variables

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

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

UFO 데이터 세트에는 위치 데이터와 만남의 유형을 포함한 많은 범주형 변수가 있고,
get_dummies를 이용해 one - hot encoding 변수를 쉽게 할 수 있음

Standardization

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

두번째 column을 var()을 이용해 열의 분산을 확인하고 numpy의 log 함수를 사용해 열을 정규화할 수 있음.

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Engineering new features

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Sarah Guido
Senior Data Scientist

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&#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`

UFO 데이터 세트에는 피쳐 엔지니어링의 훌륭한 후보인 몇 가지 필드가 있습니다. 날짜 필드에서, 우리는 목격의 달을 알고 싶을지도 모릅니다. 시간 길이 필드에서 추출해야 하는 시간(분)입니다. 마지막으로 설명 필드에는 조준에 대한 텍스트 설명이 포함되어 있습니다. 그 텍스트를 벡터화해서 우리가 그것으로부터 무엇을 배울 수 있는지 보는 것은 흥미로울 것이다. 날짜 추출을 위해 기억해야 할 몇 가지 중요한 코드는 월과 시와 같은 속성을 사용하여 필요한 날짜의 조각을 가져오는 것입니다. 정규식은 텍스트에서 숫자를 추출하는 데 도움이 되며 그룹을 사용하여 결과를 반환할 수 있습니다. 마지막으로, skikit - learn과 tf - idf vectorizer는 텍스트 필드를 벡터화합니다.

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Feature selection and modeling

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Sarah Guido
Senior Data Scientist

datacamp

Feature selection and modeling

- Redundant features
- Text vector

이 데이터를 모델링하기 전에 형상을 약간 선택해야 합니다. 중복 기능을 제거하고자 하는 경우 이 데이터셋에는 원래 형태와 기능 엔지니어링으로 인해 중복 기능을 제거할 수 있는 몇 가지 후보가 있습니다. 우리는 또한 단어를 검사하고 제거할 수 있는 텍스트 벡터를 가지고 있다. 모델링에 관한 한, 많은 연습이 있었고, 이제 사전 처리 작업의 결과를 볼 수 있습니다.

Final thoughts

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

마지막으로, 전처리 및 모델링은 종종 반복적인 작업 방식이며, 모델의 성능을 향상시키는 이상적인 기능 구성을 찾는 데 몇 번의 시도가 필요할 수 있습니다. 또한 작업 중인 데이터 세트에 대해 매우 잘 알고 구축하려는 모델을 잘 이해하는 데 도움이 됩니다.

Let's practice!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Congratulations!

PREPROCESSING FOR MACHINE LEARNING IN PYTHON

Sarah Guido
Senior Data Scientist