

Утечка данных — ситуация, когда во время обучения использована информация, которая не будет доступна во время прогнозирования.

Преобразование категориальных переменных

OrdinalEncoder (англ. «порядковый кодировщик») переводит строковые категории в числа от 0 до $n - 1$, где n — количество уникальных категорий.

```
from sklearn.preprocessing import OrdinalEncoder

ordinal_encoder = OrdinalEncoder(
    categories=[
        ['плохо', 'хорошо'],
    ], # задаёт порядок категорий
    handle_unknown='use_encoded_value', # неизвестной категории будет присвоено
    значение unknown_value
    unknown_value=np.nan # задаёт значение для неизвестных категорий
)

example_train['ordinal'] = ordinal_encoder.fit_transform(example_train)
example_test['ordinal'] = ordinal_encoder.transform(example_test)

print(example_test)
```

	Оценка комфортности покупки билета онлайн	ordinal
0	хорошо	1.0
1	плохо	0.0
2		NaN
3	плохо	0.0

Преобразование целевого признака

LabelEncoder переводит строковые значения целевого признака в порядковые числа.

```
from sklearn.preprocessing import LabelEncoder
import pandas as pd

# создание данных
y_train = pd.Series(['яблоко', 'груша', 'груша', 'груша', 'яблоко'])
y_test = pd.Series(['груша', 'груша', 'яблоко'])

# экземпляр класса LabelEncoder для кодирования целевого признака
label_encoder = LabelEncoder()

y_train = label_encoder.fit_transform(y_train)
y_test = label_encoder.transform(y_test)

print(y_test)
```

```
[0 0 1]
```

Создание полиномиальных признаков

```
from sklearn.preprocessing import PolynomialFeatures
import pandas as pd
import numpy as np

# создание данных
X = np.array([0, 1, 1, 2, 3, 4]).reshape(-1, 1)

# создание и обучение экземпляра объекта для полиномизации до второй степени
poly = PolynomialFeatures(degree=2).fit(X)

# преобразование признаков
X_poly = poly.transform(X)

# вывод результата в датафрейм
pd.DataFrame(X_poly, columns=['feature^0', 'feature^1', 'feature^2'])
```

	feature^0	feature^1	feature^2
0	1.0	0.0	0.0
1	1.0	1.0	1.0
2	1.0	1.0	1.0
3	1.0	2.0	4.0
4	1.0	3.0	9.0
5	1.0	4.0	16.0