01 - Линейная регрессия - ЛР - Insurance (Medical Cost Personal Datasets)

Необходимые нам библиотеки:

```
import numpy as np #для матричных вычислений
import pandas as pd #для анализа и предобработки данных
import matplotlib.pyplot as plt #для визуализации
import seaborn as sns #для визуализации

from sklearn import linear_model #линейные модели
from sklearn import metrics #метрики
from sklearn import preprocessing #предобработка
from sklearn.model_selection import train_test_split #разделение выборки
```

Прочитаем данные:

```
data = pd.read_csv('data/insurance.csv')
data.head()
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Рис. 1: bda90ae1ceab08737c47ab861af2941e.png

Итак, набор данных содержит следующие столбцы:

- *age* — возраст страхователя;
- *sex* — пол;
- *bmi* — индекс массы тела (), в идеале — от 18.5 до 24.9;
- *children* — количество детей, охваченных медицинской страховкой;
- *smoker* — является ли человек курящим;
- *region* — район проживания в США (северо-восток, юго-восток, северо-запад, юго-запад);
- *charges* (целевой признак) — индивидуальные медицинские расходы, оплачиваемые медицинской страховкой.

Размер таблицы:

```
print(data.shape)
# (1338, 7)
```

Выведем информацию о пропусках, так как наличие пропусков не позволит нам построить модель линейной регрессии:

```
display(data.isnull().sum())
```

Пропуски в таблице отсутствуют.

Посмотрим на типы данных:

Наши данные содержат несколько типов признаков: * *age*, *bmi*, *children* — числовые признаки; * *sex*, *smoker* — бинарные категориальные переменные (две категории); * *region* — множественные категориальные переменные (несколько категорий); * *charges* — числовой целевой признак.

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Рис. 2: 881fe5a900bea093fb7c60f07ad02bd3.png

```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

Рис. 3: 6b360058485fcbc5c5016fd578a09023.png

Прежде чем приступать к этапу подготовки данных для модели, вы можете произвести небольшое исследование зависимостей в данных, например построить следующие графики и диаграммы:

- гистограммы/коробчатые диаграммы числовых признаков;
- столбчатые диаграммы медианных медицинских расходов в зависимости от категориальных признаков;
- диаграммы рассеяния зависимости целевого признака от других числовых в разрезе категориальных (обратите особенное внимание на зависимость медицинских расходов от признака курения).

Мы знаем, что модель линейной регрессии не умеет работать с категориальными признаками, поэтому категории необходимо перекодировать.

Кодировку будем совершать по следующему принципу:

- *smoker* — переведём в бинарные значения (0 — некурящий, 1 — курящий);
- *sex* — аналогично (0 — *female*, 1 — *male*);
- *region* — используем *OneHot*-кодирование (воспользуемся функцией get_dummies).

```
#кодируем бинарные категориальные признаки
data['smoker'] = data['smoker'].apply(lambda x: 0 if x == 'no' else 1)
data['sex'] = data['sex'].apply(lambda x: 0 if x == 'female' else 1)
#оставшиеся категориальные признаки кодируем с помощью OneHot
data = pd.get_dummies(data)
data.head()
```

	age	sex	bmi	children	smoker	charges	region_northeast	region_northwest	region_southeast	region_southwest
0	19	0	27.900	0	1	16884.92400	0	0	0	1
1	18	1	33.770	1	0	1725.55230	0	0	1	0
2	28	1	33.000	3	0	4449.46200	0	0	1	0
3	33	1	22.705	0	0	21984.47061	0	1	0	0
4	32	1	28.880	0	0	3866.85520	0	1	0	0

Итак, мы получили перекодированную таблицу, в которой все признаки являются числовыми.

Выделим факторы и целевой признак в отдельные таблицы:

```
features = data.drop('charges', axis=1).columns
X, y = data[features], data['charges']
```

Теперь мы можем начинать работу над моделью.

Задания

1 Прежде чем переходить к этапу моделирования, нам необходимо позаботиться о создании выборки для тестирования модели.

Разделите набор данных на тренировочную и тестовую выборки в соотношении 80/20. Воспользуйтесь функцией `train_test_split`.

В качестве значения параметра `random_state` укажите число 42.

Чему равно количество наблюдений в тестовом наборе данных?

2 Обучите модель линейной регрессии аналитическим методом (*LinearRegression*) на тренировочном наборе данных. Все параметры оставьте по умолчанию.

Чему равен свободный член (*intercept*) обученной модели? Ответ округлите до сотых.

С помощью модели, полученной в предыдущем задании, сделайте предсказание на тренировочной и тестовой выборке. Рассчитайте следующие три метрики: R^2, MAE, $MAPE$. Не забудьте привести значение $MAPE$ к процентам.

Для удобства проверки значение R^2 округлите до трёх знаков после точки-разделителя, а значения MAE и $MAPE$ — до целого числа.

Чему равны значения метрик на тренировочной и тестовой выборках?

3 Постройте диаграмму *boxplot* для визуализации ошибок модели линейной регрессии на тренировочной и тестовой выборках. В качестве ошибки возьмите разницу между истинным ответом и предсказанием: $y - \hat{y}$ (без модуля).

Выберите верные ответы: **A** Разброс ошибок на тестовой выборке больше, чем на тренировочной. **B** Разброс ошибок на тренировочной выборке больше, чем на тестовой. **C** Медианная ошибка на тренировочной и тестовой выборках отрицательная (меньше 0). **D** Медианная ошибка на тренировочной и тестовой выборках положительная (больше 0)

4 Нормализуйте тренировочную и тестовую выборки с помощью *min-max*-нормализации (*MinMaxScaler*). Расчёт параметров нормализации (`fit`) произведите на тренировочной выборке.

Примечание. *Min-max*-нормализация не искажает изначальный вид бинарных категориальных признаков, в отличие от стандартизации.

На нормализованных данных сгенерируйте полиномиальные признаки степени 2. Воспользуйтесь классом `PolynomialFeatures` из библиотеки *sklearn*. Значение параметра `include_bias` выставите на `False`.

Чему равно результирующее количество столбцов?

5 Обучите модель линейной регрессии на полиномиальных признаках.

Чему равно значение метрики R^2 на **тестовой** выборке?

Значение R^2 округлите до трёх знаков после запятой.

6 Выведите значения коэффициентов полученной модели. Посмотрите на степени коэффициентов.

Какой вывод можно сделать? - Значения коэффициентов очень высокие, модель неустойчива, необходима регуляризация. - Значения коэффициентов приемлемые, модель устойчива, регуляризация не нужна.

7 Постройте линейную регрессию с *L1*-регуляризацией (*Lasso*) на полиномиальных признаках. В качестве параметра `alpha` используйте значение по умолчанию, параметр `max_iter` установите в значение 2000.

Чему равны метрики R^2, MAE и $MAPE$ на **тестовой** выборке?

Значение R^2 округлите до трёх знаков после запятой, а значения MAE и $MAPE$ до целого числа.