# Решение задачи регрессии с помощью методов машинного обучения

Голиков М. О.

«ОРЛОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ И. С. ТУРГЕНЕВА»

19 декабря 2020

# Содержание

# Описание исходных данных

В данной работе используются **данные** о страховых выплатах медицинскому персоналу.

| | age | sex | bmi | children | smoker | region | charges |
|---|---|---|---|---|---|---|---|
| 2 | 19 | female | 27.9 | 0 | yes | southwest | 16884.924 |
| 3 | 18 | male | 33.77 | 1 | no | southeast | 1725.5523 |
| 4 | 28 | male | 33 | 3 | no | southeast | 4449.462 |
| 5 | 33 | male | 22.705 | 0 | no | northwest | 21984.47061 |
| 6 | 32 | male | 28.88 | 0 | no | northwest | 3866.8552 |
| 7 | 31 | female | 25.74 | 0 | no | southeast | 3756.6216 |
| 8 | 46 | female | 33.44 | 1 | no | southeast | 8240.5896 |
| 9 | 37 | female | 27.74 | 3 | no | northwest | 7281.5056 |

Рисунок: Пример данных

# Описание исходных данных

Описание признаков:

- age - возраст бенефициара (лица, получающего страховые выплаты);
- sex - пол бенефициара;
- bmi - индексы массы тела;
- children - количество детей бенефициара, на которых распространяется страхование / количество иждевенцов бенефициара;
- smoker - является ли бенефициар курильщиком;
- region - жилой район бенефициара в США: северо-восток, юго-восток, юго-запад, северо-запад;
- charges - индивидуальные медицинские расходы, выставленные на счет медицинского страхования.

# Анализ данных

## Гистограммы признаков sex и region


Рисунок: Гистограмма признака sex

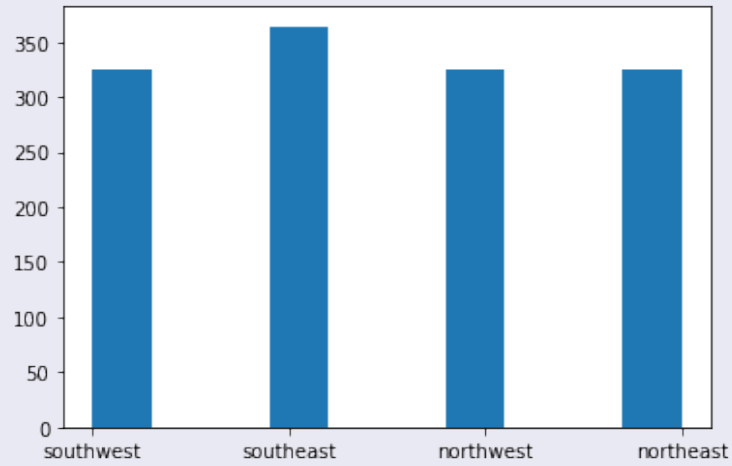



Рисунок: Гистограмма признака region

# Анализ данных

## Гистограммы признаков children и smoker

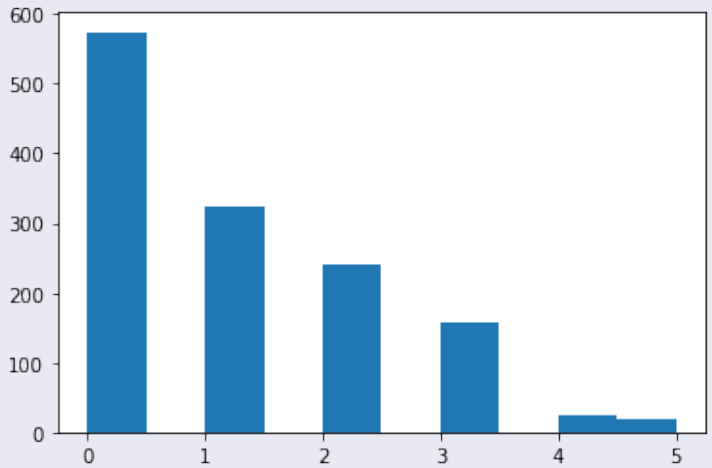



Рисунок: Гистограмма признака children

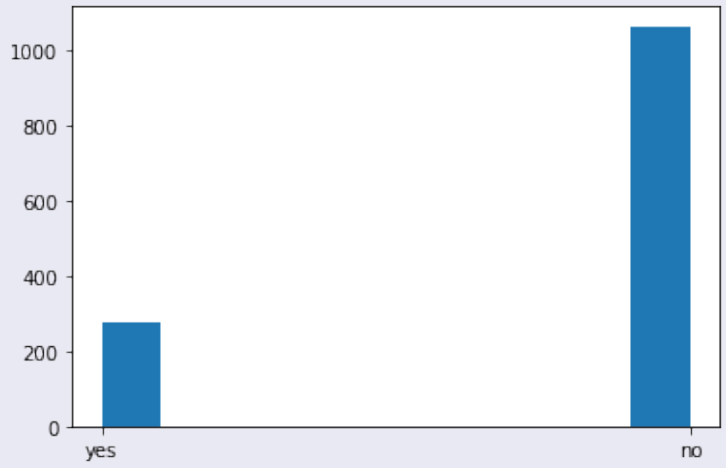



Рисунок: Гистограмма признака smoker

# Анализ данных

С помощью гистограмм удалось выяснить, что в категориальных признаках нет категорий, у которых была бы значительная количественная разница по отношению к другим значениям.

# Анализ данных

## Диаграммы размаха признаков sex и region

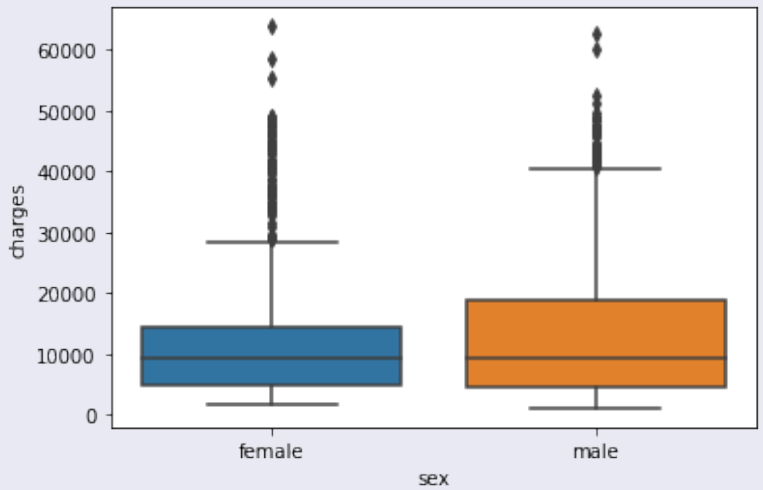



Рисунок: Гистограмма признака sex

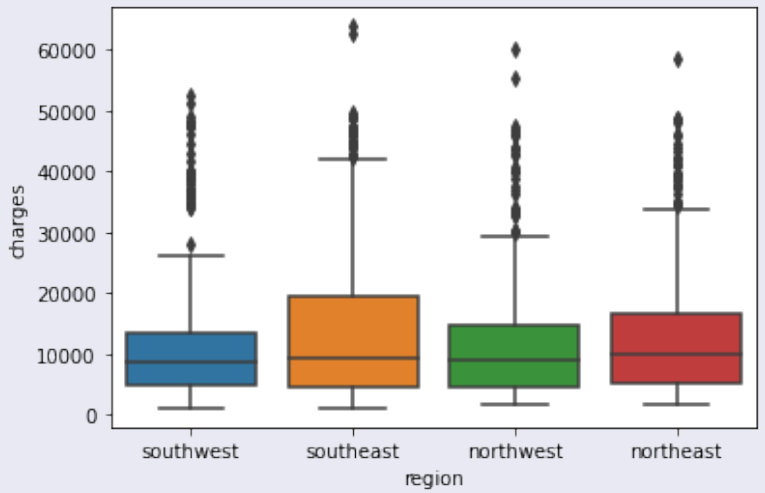



Рисунок: Гистограмма признака region

# Анализ данных

## Диаграммы размаха признаков children и smoker

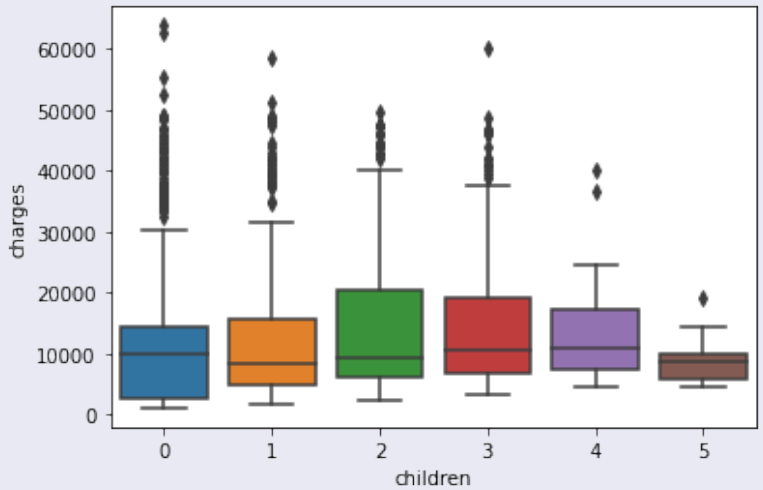



Рисунок: Гистограмма признака children

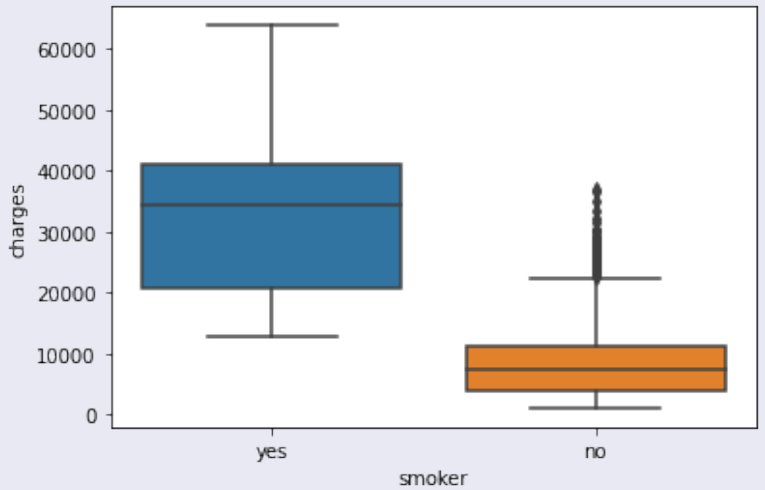



Рисунок: Гистограмма признака smoker

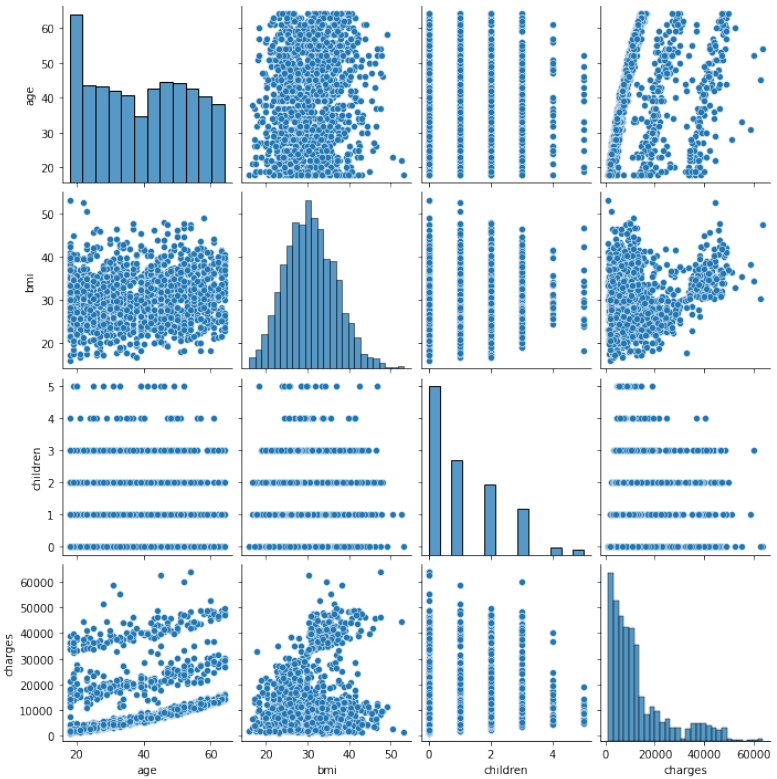
# Анализ данных

С помощью диаграмм размаха удалось выяснить, что в между категориями имеются статистические различия.

# Анализ данных

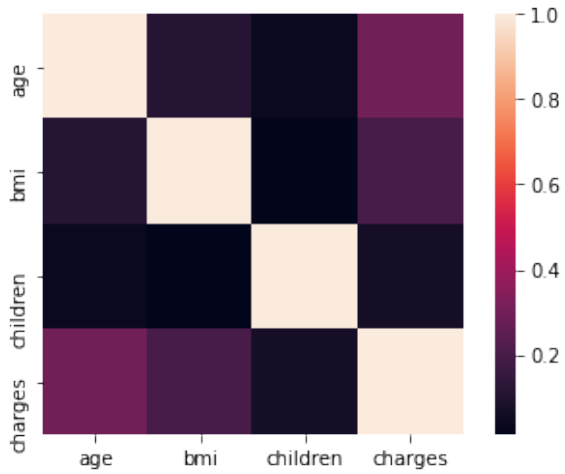
Справа можно увидеть диаграмму рассеяния и распределения признаков. Можно заметить, что между признаками не наблюдается линейной зависимости.

Стоит обратить внимание на распределение признака charges, которое сильно скошено в левую сторону.



# Анализ данных

На рисунке справа представлена визуализация матрицы корреляции. На основе этих данных можно сделать вывод, что между вещественными признаками не наблюдается линейной зависимости, а также отсутсвует мультиколлинеарность.



# Моделирование

В качестве базиса используется линейная регрессия.
Ввиду отсутствия мультиколлениарности и небольшого количества признаков применение таких алгоритмов как LASSO, Ridge и Elastik-Net кажется нецелесообразным.
Можно использовать случайный лес, т. к. он хорошо умеет восстанавливать нелинейные зависимости, но нужно помнить о том, что случайный лес не способен экстраполировать.
Для решения данной задачи также можно использовать метод опорного вектора регрессии.

# Моделирование

В работе используются следующие алгоритмы:

- линейная регрессия;
- опорный вектор регрессии;
- случайный лес.

Для подбора наиболее оптимальных параметров используется GridSearchCV.

Для оценки алгоритмов используется Root Mean Square Error (RMSE) – среднеквадратическое отклонение ошибки. Данная оценка позволит оценить разброс ошибки.

# Моделирование

В таблице ниже можно увидеть значения RMSE для используемых алгоритмов.

Таблица: Оценка ошибки алгоритмов

| Алгоритм | RMSE |
|---|---|
| Линейная регрессия | 0,434 |
| Опорный вектор регрессии | 0,443 |
| Случайный лес | 0,371 |

# Моделирование

Мы видим, что у случайный леса наилучшее значение RMSE. Однако, нужно помнить что случайный лес не способен экстраполировать.
Можно использовать композицию алгоритмов. Применив композицию на основе голосования получаем значение RMSE равное 0,396.