Тимошинов Егор Борисович, группа 16

Лабораторная работа

Логистическая регрессия

Цель работы

Изучение методов построения моделей логистической регрессии и оценки их качества с помощью ROC-анализа.

Теоретические сведения

Логистическая регрессия используется для моделирования зависимости бинарной зависимой переменной от независимых переменных. В отличие от линейной регрессии, логистическая регрессия предсказывает вероятность того, что зависимая переменная примет значение 1.

Модель логистической регрессии имеет вид:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

где $P(Y = 1|X)$ — вероятность того, что зависимая переменная Y равна 1 при заданных значениях независимых переменных X, β_0 — свободный член, β_1 — коэффициент регрессии.

Задание

1. Построить модель логистической регрессии для предсказания вероятности заболевания в зависимости от возраста.
2. Построить график зависимости вероятности заболевания от возраста.
3. Построить ROC-кривую и вычислить значение AUC.
4. Построить матрицу ошибок и проанализировать качество классификации.
5. Проанализировать распределение предсказанных вероятностей.

Результаты выполнения задания

Исходные данные

Для анализа использованы данные о 100 наблюдениях. Зависимая переменная — наличие заболевания 0—*нетзаболевания*, 1—*естьзаболевание*, независимая переменная — возраст.

Статистика по данным:

- Количество случаев заболевания: 60
- Количество случаев без заболевания: 40
- Средний возраст: 54.58 лет
- Минимальный возраст: 26 лет
- Максимальный возраст: 84 лет

Первые 20 наблюдений:

Возраст	Заболевание	Предсказанная вероятность	Предсказанный класс
26	0	0.0796	0
26	0	0.0796	0
26	1	0.0796	0
26	0	0.0796	0
27	0	0.0882	0
27	0	0.0882	0
28	0	0.0977	0
28	0	0.0977	0
28	0	0.0977	0
30	0	0.1194	0
31	0	0.1317	0
31	0	0.1317	0
32	0	0.1452	0
32	0	0.1452	0
32	0	0.1452	0
33	0	0.1597	0
33	0	0.1597	0
34	0	0.1754	0
35	1	0.1922	0
35	1	0.1922	0

Результаты моделирования

Коэффициенты модели логистической регрессии:

- Свободный член β_0: -5.3742
- Коэффициент при возрасте β_1: 0.1125

Уравнение модели:

$$P(\text{заболевание}|\text{возраст}) = \frac{1}{1+e^{-(-5.3742+0.1125 \cdot \text{возраст})}}$$

График зависимости вероятности заболевания от возраста

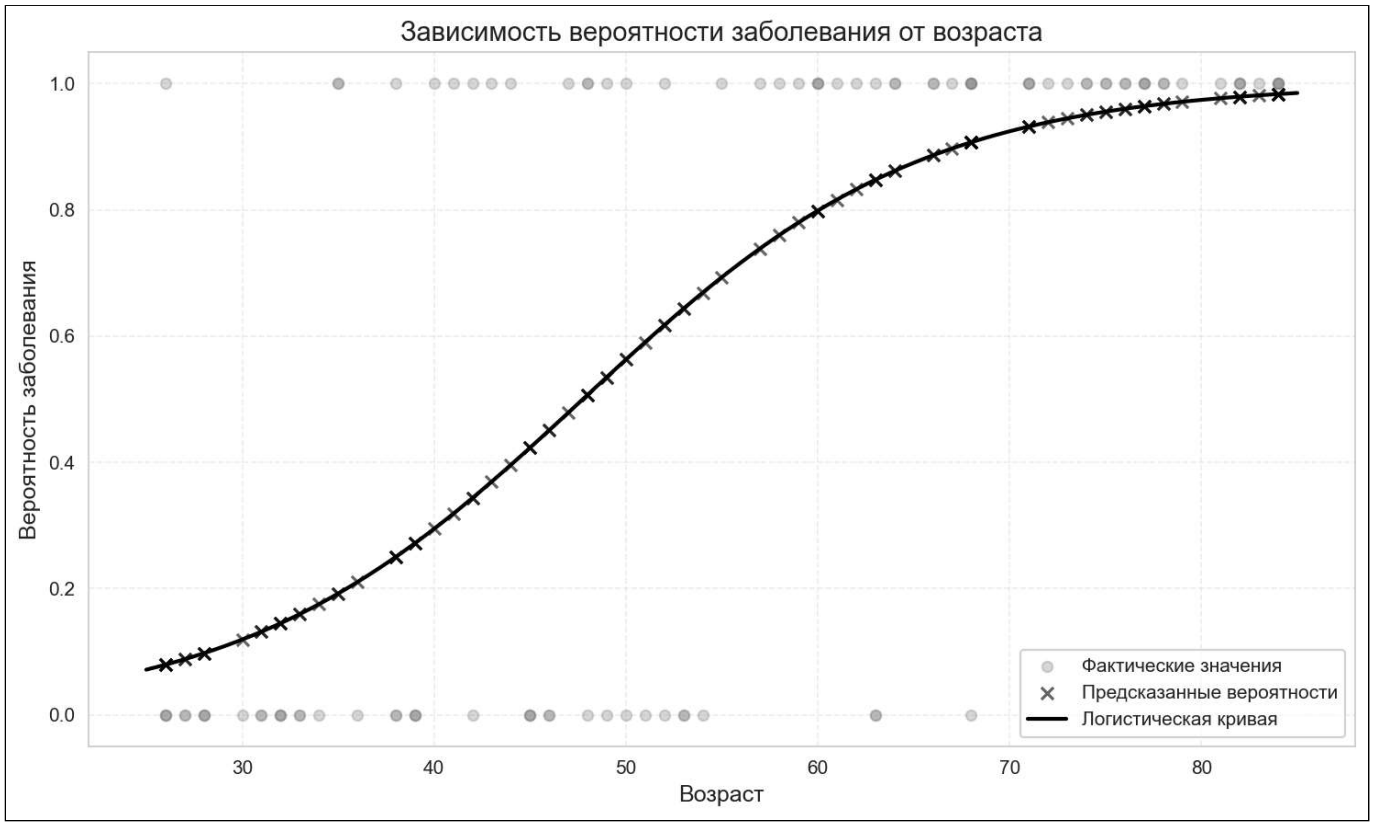

Рисунок 1. Зависимость вероятности заболевания от возраста

На графике видно, что вероятность заболевания увеличивается с возрастом. Логистическая кривая имеет S-образную форму, что характерно для логистической регрессии.

ROC-кривая

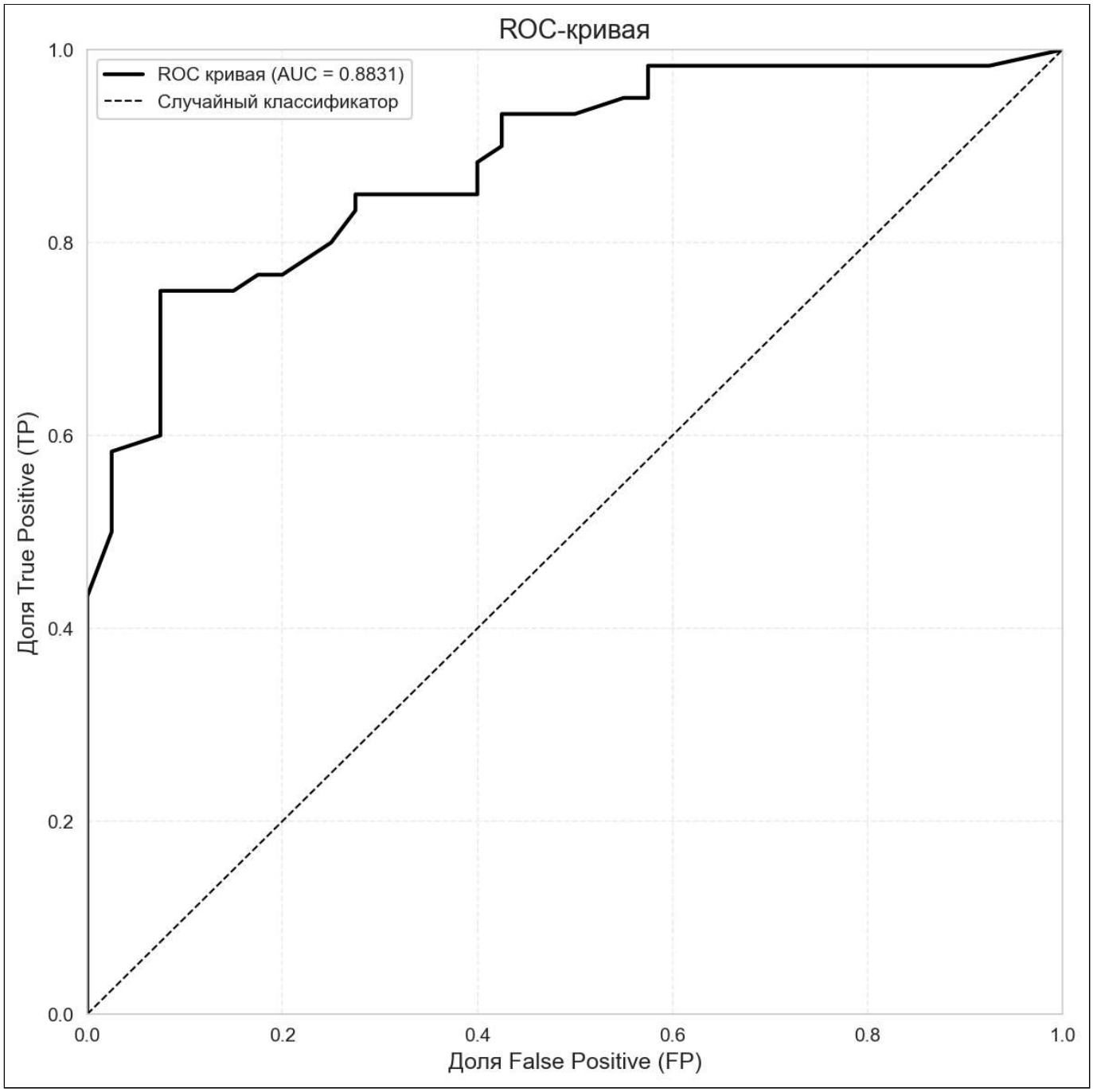

Рисунок 2. ROC-кривая

Площадь под ROC-кривой *AUC* составляет 0.8831. Значение AUC близкое к 1 указывает на хорошее качество модели. Значение 0.5 соответствует случайному классификатору.

Матрица ошибок

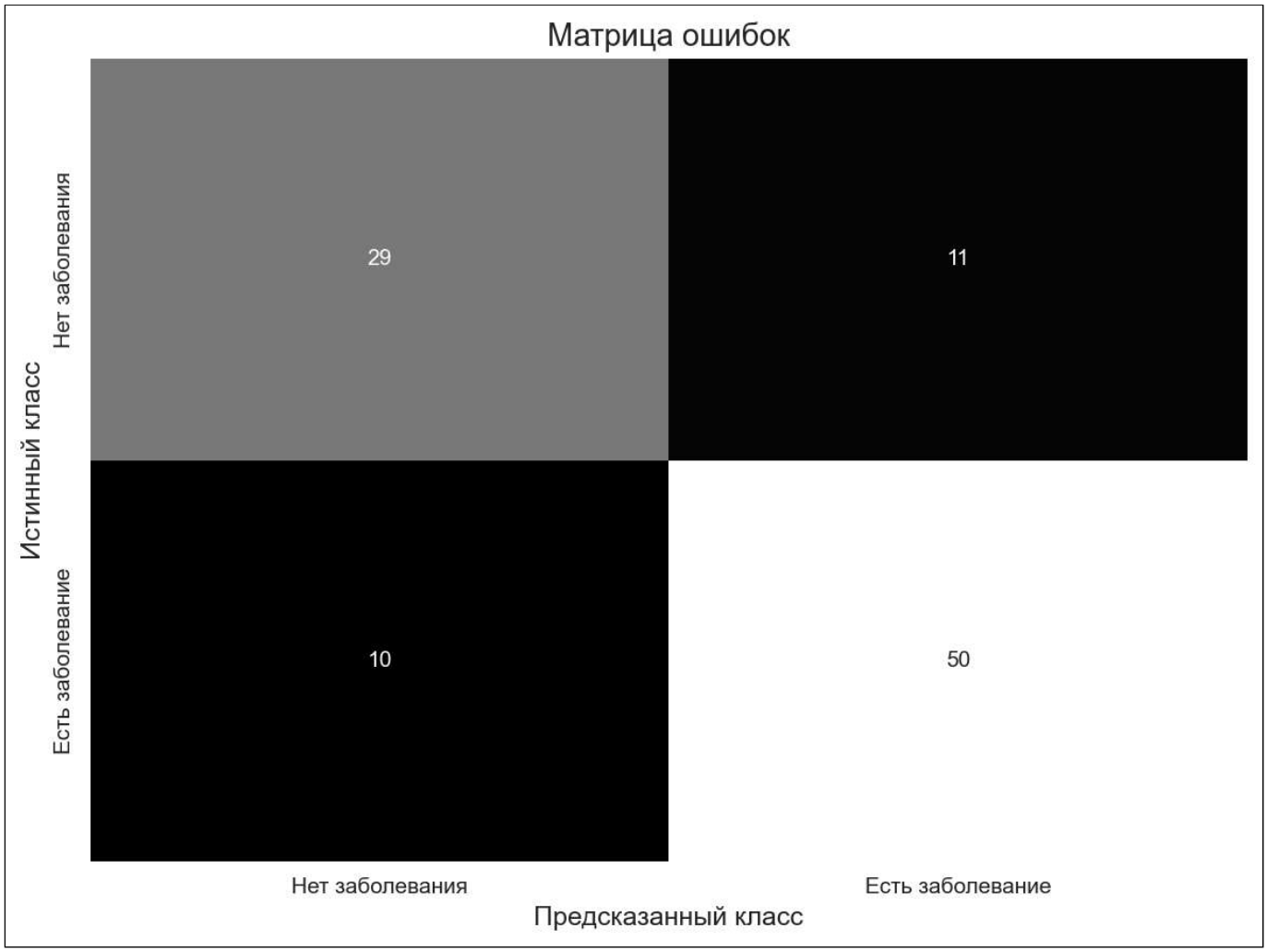

Рисунок 3. Матрица ошибок

Матрица ошибок показывает количество правильных и неправильных классификаций:

- True Negative TN: 29 — правильно предсказано отсутствие заболевания
- False Positive FP: 11 — ошибочно предсказано наличие заболевания
- False Negative FN: 10 — ошибочно предсказано отсутствие заболевания
- True Positive TP: 50 — правильно предсказано наличие заболевания

Распределение предсказанных вероятностей

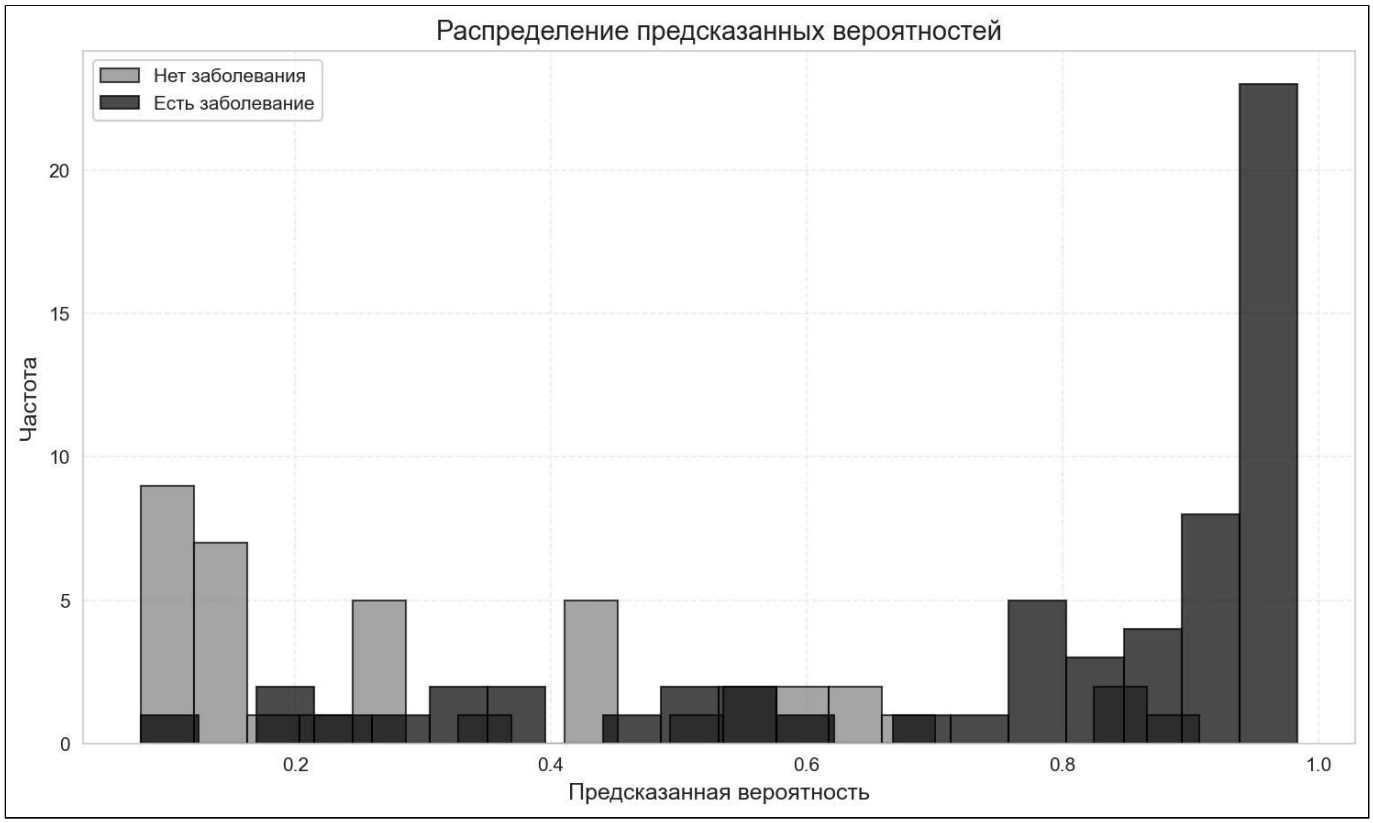

Рисунок 4. Распределение предсказанных вероятностей

Гистограмма показывает распределение предсказанных вероятностей для случаев с заболеванием и без него. Хорошая модель должна давать низкие вероятности для случаев без заболевания и высокие — для случаев с заболеванием.

Оценка качества модели

Метрики качества классификации:

- Точность *Accuracy*: 0.7900
- Precision *дляклассa*1: 0.8197
- Recall *дляклассa*1: 0.8333
- F1-score *дляклассa*1: 0.8264

Выводы

В ходе выполнения лабораторной работы была построена модель логистической регрессии для предсказания вероятности заболевания в зависимости от возраста. Модель показала хорошее качество $AUC = 0.8831$, что указывает на наличие значимой связи между возрастом и вероятностью заболевания. ROC-кривая демонстрирует, что модель лучше случайного классификатора. Матрица ошибок и распределение вероятностей подтверждают адекватность построенной модели.