ЛЕКЦИЯ 13. Логическая регрессия

Основы логистической регрессии

Линейная регрессия используется для моделирования линейных зависимостей между непрерывной выходной переменной и набором входных переменных. При анализе данных часто встречаются задачи, где выходная переменная является категориальной и тогда использование линейной регрессии затруднено. Поэтому при поиске связей между набором входных переменных и категориальной выходной переменной получила распространение логистическая регрессия. Рассмотрим применение логистической регрессии для случая бинарной выходной переменной (переменной, которая может принимать только два значения), хотя можно использовать данный метод и в случае, когда выходная переменная принимает более чем два значения [3].

1.7.1. Простой пример логистической регрессии

Предположим, что врача интересует зависимость между возрастом пациента и наличием (1) или отсутствием (0) некоторого заболевания. Данные, собранные по 20 пациентам, представлены в табл. 1.9, а соответствующий график на рис. 1.12. Таким образом, в задаче используется бинарная выходная переменная y, которая может принимать только два значения: 0 и 1. Иногда такие переменные называют дихотомическими.

Таблица 1.9. Данные о пациентах

№ пациента	Возраст, x	Наличие заболевания, y
1	25	0
2	29	0
3	30	0
4	31	0
5	32	0
6	41	0
7	41	0
8	42	0
9	44	1
10	49	1
11	50	0
12	59	1
13	60	0
14	62	0

15	68	1
16	72	0
17	79	1
18	80	0
19	81	1
20	84	1

На рисунке сплошной линией представлена прямая простой линейной регрессии, построенная для данных из табл. 1.12, а пунктиром кривая логистической регрессии. Также для обеих кривых показана ошибка оценивания для пациента №11 (x = 50, y = 0). Линия логистической регрессии, в отличие от линейной, не является прямой.

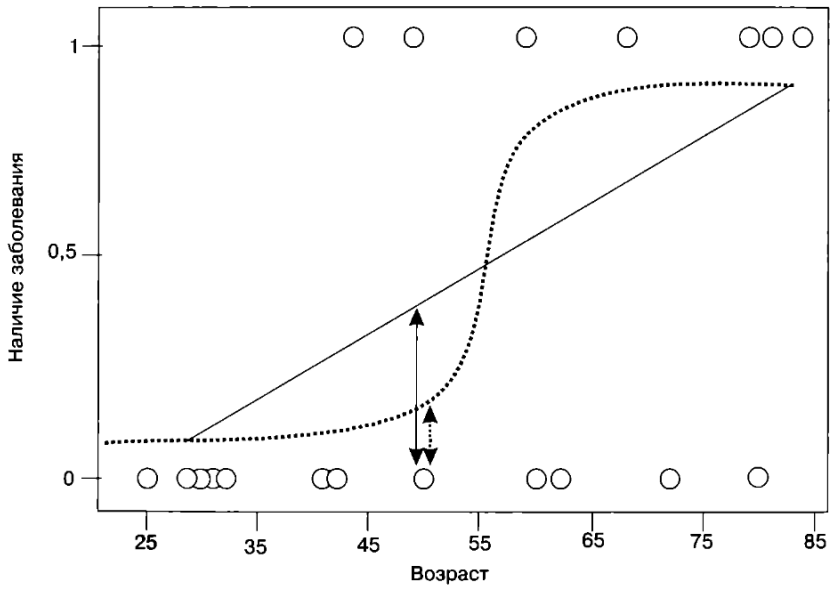

Рисунок 1.12. Диаграмма «Возраст – заболевание», линия регрессии и кривая логистической регрессии

Рассмотрим ошибки оценивания, полученные для пациента №11. Расстояние между точкой данных для пациента №11 и линией регрессии показано сплошной вертикальной стрелкой, а для кривой логистической регрессии пунктирной. Видно, что расстояние будет больше для линейной регрессии, а это означает, что она дает худшую оценку выходной переменной, чем логистическая. Это утверждение также является истинным для большинства других пациентов.

Построение линии логистической регрессии

Введем в рассмотрение условное среднее $E(y|x)$ значений выходной переменной y для заданного значения x входной переменной X. $E(y|x)$ представляет собой ожидаемое значение выходной переменной при заданном значении входной. Напомним, что выходная переменная в линейной регрессии – это случайная переменная, определяемая как $y = \beta_0 + \beta_1 x_1 + \varepsilon$. Поскольку

ошибка ε имеет нулевое среднее, для линейной регрессии мы получим, что $E(y|x) = \beta_0 + \beta_1 x_1$ (так же как и в линейной регрессии, буквой b будем обозначать коэффициенты уравнения регрессии, а β – параметры соответствующей модели) [4].

Для краткости введем обозначение $E(y|x) = \rho(x)$. Условное среднее для логистической регрессии имеет вид:

$$\rho(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (1.8)$$

Функцию, описываемую уравнением (1.8), называют логистической, а соответствующие кривые – сигмоидами, поскольку они имеют характерную S-образную форму. Эта функция определена на бесконечности и изменяется в диапазоне от 0 до 1. Диапазон изменения $\rho(x)$ также будет от 0 до 1, поэтому данную функцию можно интерпретировать как вероятность того, что выходная переменная приобрела значение 1 (заболевание имеет место), а $1 - \rho(x)$ – как вероятность появления значения 0 (заболевание отсутствует).

Как уже говорилось при обсуждении модели линейной регрессии, ошибка ε является нормально распределенной случайной величиной с нулевым средним и постоянной дисперсией. Предположения, используемые для логистической регрессии, несколько отличаются. Выходная переменная является бинарной, и принятие выходной переменной одного из двух возможных значений называется исходом [2].

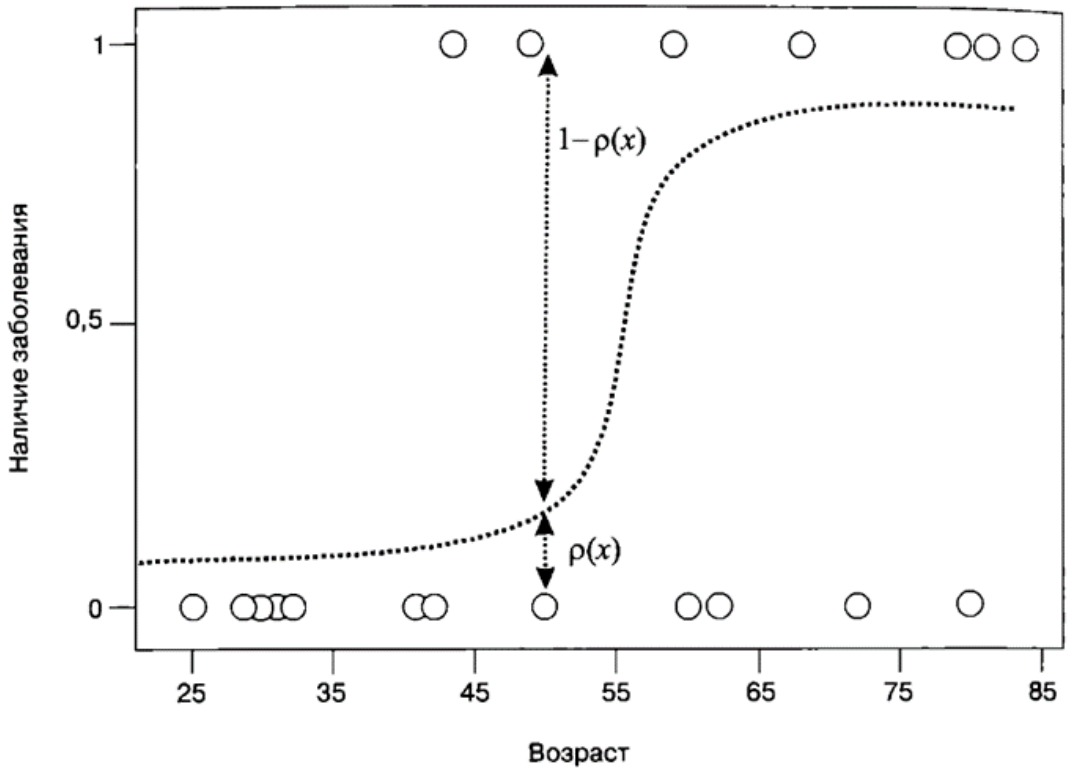

Рисунок 1.13. Геометрическая интерпретация вероятности исхода заболевания

Исход – явление, показатель или признак, который служит объектом исследования. Например, при проведении клинических испытаний в медицине вероятность исхода служит критерием оценки эффективности лечебного или профилактического воздействия [4].

Если предположить, что принятие выходной переменной y значения 1 рассматривается как успех, а значения 0 – как неуспех, то $\rho(x)$ можно интерпретировать как вероятность успеха, а $1-\rho(x)$ – неуспеха. Данная ситуация поясняется на рис. 1.13.

В логистической регрессии используется преобразование вида:

$$g(x) = \ln\frac{\rho(x)}{1-\rho(x)} = \beta_0 + \beta_1 x$$

Оно называется логит-преобразованием и обладает такими полезными свойствами, как линейность, непрерывность и определенность на бесконечности.

Оценки максимального правдоподобия

Одним из наиболее привлекательных свойств линейной регрессии является то, что коэффициенты регрессии могут быть получены с помощью метода наименьших квадратов. Для оценки коэффициентов логистической регрессии таких решений не существует. Поэтому в ней коэффициенты оцениваются на основе метода максимального правдоподобия, который позволяет найти такие значения коэффициентов, для которых вероятность появления максимальна [3].

Введем в рассмотрение функцию правдоподобия $l(\beta|x)$. Она определяет вероятность появления значений параметров $\beta = \beta_1, \beta_2 \dots \beta_\mu$ для заданного значения x. Задача заключается в поиске таких значений этих параметров, которые максимизируют функцию правдоподобия: строятся оценки максимального правдоподобия, для которых значения параметров являются наиболее подходящими для наблюдаемых данных.

Вероятность того, что выходная переменная y приобретет значение 1 для заданного значения x (вероятность успеха), будет $\rho(x) = P(y=1|x)$, а вероятность того, что $y=0$ при заданном x, будет $1-\rho(x) = P(y=0|x)$. Таким образом, поскольку y_i или 1, вклад i-го наблюдения может быть выражен как $[\rho(x_i)]^{y_i} \times [1-\rho(x_i)]^{(1-y_i)}$. Предположение, что наблюдения являются независимыми, позволяет представить функцию правдоподобия как произведение двух отдельных членов [1]:

$$l(\beta|x) = \prod_{i=1}^{n} [\rho(x_i)]^{y_i} \times [1-\rho(x_i)]^{(1-y_i)}$$

В вычислительном плане более удобна логарифмическая функция правдоподобия $L(\beta|x) = \ln[l(\beta|x)]$

$$L(\beta|x) = \ln[l(\beta|x)] = \sum_{i=1}^{n}\{y_i \ln[\rho(x_i)] + (1 - y_i)\ln[1 - \rho(x_i)]\} \qquad (1.9)$$

Оценки максимального правдоподобия могут быть найдены путем дифференцирования $L(\beta|x)$ относительно каждого параметра и приравниванием результирующих выражений к 0.

Проверим результаты логистической регрессии для данных из табл. 1.9. Коэффициенты, то есть оценки максимального правдоподобия неизвестных параметров β_0 и β_1 определятся как $\beta_0 = -4{,}372$, а $\beta_1 = 0{,}067$. С учетом уравнения (1.8) можно записать:

$$\hat{\rho}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{-4{,}372+0{,}067(\text{возраст})}}{1 + e^{-4{,}372+0{,}067(\text{возраст})}}$$

где $\hat{g}(x) = -4{,}372 + 0{,}067x$ есть логит-преобразование.

Эти уравнения могут использоваться, чтобы оценить вероятность наличия заболевания у пациентов определённого возраста. Например, для пациента в возрасте 50 лет имеем:

$\hat{g}(x) = -4{,}372 + 0{,}067 \times 50 = -1{,}022;$

$\hat{\rho}(x) = \frac{e^{\hat{g}(x)}}{1+e^{\hat{g}(x)}} = \frac{e^{-1{,}022}}{1+e^{-1{,}022}} = 0{,}26.$

В итоге вероятность того, что пациент 50 лет страдает заболеванием, составляет 26%. Соответственно, вероятность отсутствия заболевания будет 100-26=74 %. Если провести такую же оценку для пациента в возрасте 72 лет, то можно увидеть, что вероятность наличия заболевания составит 61%, а его – отсутствия 39%.

Значимость входных переменных

Напомним, что в простой линейной регрессии модель считалась значимой, если средний квадрат значении оценок регрессии был больше, чем средний квадрат ошибки оценивания. Средний квадрат регрессии представляет собой меру улучшения оценки выходной переменной, если для оценивания мы используем не среднее значение, а входную переменную. Если входная переменная является «полезной» для оценивания значения выходной, то средний квадрат регрессии будет больше и статистика $F = Q_R/Q$ также будет больше. Тогда соответствующую линейную регрессионную модель можно будет рассматривать как значимую [2].

Значимость коэффициентов логистической регрессии определяется аналогично. В сущности, мы проверяем, обеспечивает ли использование в модели определенной входной переменной лучшую оценку выходной переменной.

Введем понятие насыщенной модели, то есть модели, в которой количество входных переменных равно числу наблюдений данных. Очевидно, что такая модель будет предсказывать значения выходной переменной с абсолютной точностью. Затем мы можем посмотреть на наблюдаемые значения выходной переменной, которые были предсказаны с помощью насыщенной модели. Для сравнения оценок, полученных с помощью обычной модели и насыщенной модели, введем понятие отклонения D [2]:

$$D = -2\ln\frac{l(\beta|x)}{l_{\text{нас}}(\beta|x)} \tag{1.10}$$

Здесь мы имеем опошление двух значений функции правдоподобия, поэтому проверка результирующей гипотезы называется проверкой отношения правдоподобия.

Отношение правдоподобия – отношение вероятности получить положительный результат для положительного исхода к вероятности получить положительный результат для отрицательного исхода [4].

Например, для выборки из табл. 1.9 отношение правдоподобия – это отношение вероятности обнаружить болезнь у больного к вероятности обнаружить болезнь у здорового.

Кроме этого, введем еще два понятия.

Отношение правдоподобия положительного результата – отношение вероятности получить истинноположительный результат к вероятности получить ложноположительный результат.

Отношение правдоподобия отрицательного результата теста – отношение вероятности получить истинноотрицательный результат к вероятности получить ложноотрицательный результат.

Обозначим оценку $\rho(x)$, полученную с помощью обычной модели, как $\hat{\rho}(x)$. Затем, используя уравнение (1.10), для случая логистической регрессии мы можем записать отклонение D в следующем виде [1]:

$$D = -2\sum_{i=1}^{n}\left[y_i\ln\frac{\hat{\rho}_i}{y_i} + (1-y_i)\ln\frac{1-\hat{\rho}_i}{1-y_i}\right]$$

Чтобы определить, является ли переменная значимой, нужно найти разность двух отклонений вычисленного для модели без данной входной переменной D^- и найденного для всей модели D^+, то есть

$$G = D^- - D^+ = -2\ln\left[\frac{\text{правдоподобие без переменной}}{\text{правдоподобие с переменной}}\right]$$

Введем обозначения $n_1 = \sum y_i$ и $n_0 = \sum(1-y_i)$. Тогда для случая единственной входной переменной можно записать [1]:

$$G = 2\left\{\sum_{i=1}^{n}[y_i \ln \hat{\rho}_i + (1 - y_i) \times \ln(1 - \hat{\rho}_i)] - [n_1 \ln n_1 + n_0 \ln n_0 - n \ln n]\right\}$$

Для примера из табл. 1.9 логарифмическое правдоподобие будет – 10,101, тогда:

$$G = 2\{-10{,}101 - [7\ln(7) + 13\ln(13) - 20\ln(20)]\} = 5{,}696$$

При справедливости нулевой гипотезы, состоящей в предположении $\beta_1 = 0$, статистика G имеет распределение X^2 с одной степенью свободы. Следовательно, результирующее ρ-значение для проверки данной гипотезы будет $P(X_1^2 > 5{,}696) = 0{,}017$. Благодаря весьма малому ρ-значению становится очевидно, что возраст очень значимая переменная при определении вероятности наличия заболевания.

Другим методом для проверки значимости определенной входной переменной является тест Вальда. При нулевой гипотезе $\beta_1 = 0$ отношение $Z_W = \left(\frac{b_1}{E_{\text{ст}} b_1}\right)^2$ соответствует распределению хи-квадрат с одной степенью свободы, где $E_{\text{ст}} b_1$ – стандартная ошибка оценивания коэффициента регрессии на основе наблюдаемых данных.

Напомним, что стандартная ошибка оценивания коэффициентов регрессии $E_{\text{ст}}$ использовалась ранее для оценки значимости коэффициентов линейной регрессии. В логистической регрессии она также используется для этих целей. О том, как вычисляются стандартные ошибки оценивания коэффициентов логистической регрессии $E_{\text{ст}}$, будет рассказано дальше, а пока возьмем готовые цифры. Поскольку $b_1 = 0{,}067$, а $E_{b_1} = 0{,}0322$, то $Z_W = 0{,}067^2/0{,}0322^2 = 4{,}33$ и $P(z > 4{,}33) = 0{,}038$. Это неравенство означает, что вероятность справедливости нулевой гипотезы не превышает 3,8%. Данное ρ-значение также достаточно мало, хотя и не настолько, как полученное с помощью отношения правдоподобия. Следовательно, результаты обоих тестов совпадают, и переменная возраста является статистически значимой для оценки вероятности заболевания.

Введем в рассмотрение понятие доверительного интервала.

Доверительный интервал – наиболее вероятный диапазон изменения наблюдений случайной величины. Величины, полученные в исследованиях на выборке данных, отличаются от истинных (наблюдаемых) величин вследствие влияния случайной составляющей. Так 95%-й доверительный интервал означает, что истинное значение величины с вероятность 95% лежит в пределах данного интервала. Доверительные интервалы помогают определить, соответствует ли данный диапазон значений представлениям аналитика о значимости связи между

переменными. Величина доверительного интервала характеризует степень доказательности данных, в то время как ρ-значение указывает на вероятность отклонения нулевой гипотезы [3].

Определим границы доверительных интервалов для оценок коэффициента b_1 логистической регрессии $(1 - \alpha) \times 100\%$ в следующем виде:

$$b_1 \pm zE_{\text{ст}}(b_1)$$

где z определяется с использованием t-критерия Стьюдента для двусторонней области при заданном уровне значимости α и с $n - k$ степенями свободы. В нашем примере 95% доверительный интервал для коэффициента b_1 определится так:

$$CI(b_1) = 0{,}06696 \pm 1{,}96 \times 0{,}03223 = 0{,}06696 \pm 0{,}06317 = [0{,}00379;\ 0{,}13013]$$

Поскольку 0 не входит в данный интервал, мы можем заключить, что с вероятностью $95\% b_1 \neq 0$ и, следовательно, возраст пациента является значимой переменной.

Использование логистической регрессии для решения задач классификации

Постановка задач классификации и регрессии отличается характером выходной переменной. Если выходная переменная является непрерывной, то имеет место задача регрессии, а если дискретной (метка класса) – то классификации. Как было показано выше, логистическая регрессия позволяет работать с дихотомической выходной переменной, что предполагает возможность использования этого метода для решения задач бинарной классификации. В бинарной классификации каждое наблюдение или объект должны быть отнесены к одному из двух классов (например, А и Б). Тогда с каждым исходом связано событие: объект принадлежит к классу А и объект принадлежит к классу Б. Результатом будет оценка вероятности соответствующего исхода [2].

Если в процессе анализа будет установлено, что вероятность $P(A)$ принадлежности объекта с заданным набором значений признаков (входных переменных) к А больше, чем вероятность $P(\text{Б})$ его принадлежности к классу Б, то он будет классифицирован как объект класса А. Очевидно, что поскольку события взаимоисключающие, то $P(\text{Б}) = 1 - P(A)$. Может быть задан порог вероятности, при превышении которого вероятность, связанная с определенным классом, «перевешивает» и объект относится к этому классу. В простейшем случае это может быть порог равной вероятности, то есть 0,5. Как только вероятность $P(\text{Б})$ становиться 0,51, а $P(A)$ – 0,49, объект относится к классу Б. Иногда порог определяется более сложным образом, например исходя из надежности решения. Так, решение о принадлежности объекта к определенному

классу может быть принято только тогда, когда вероятность данного события, оцененная с помощью логистической регрессии, превысит 0,7.

Бинарная классификация на основе логистической регрессии широко применяется при решении задач в медицине, технической диагностике, социальной сфере и других предметных областях.

Интерпретация модели логистической регрессии

Очень важно не только математически описать модель, но и правильно интерпретировать ее с точки зрения анализа, то есть извлечь всю необходимую информацию об исследуемых объектах и процессах.

Напомним, что в простой линейной регрессии коэффициент b_1 интерпретируется как изменение значения выходной переменной при изменении входной на 1. В логистической регрессии его интерпретация аналогична, но применительно к логистической функции. То есть коэффициент b_1 может быть интерпретирован как изменение значения логистической функции при изменении входной переменной на 1. Формально это можно записать в виде [1]:

$$b_1 = g(x-1) - g(x)$$

Рассмотрим интерпретацию коэффициента b_1 в простой логистической регрессии для трех случаев [2]:

- когда входная переменная принимает только два значения (дихотомическая входная переменная);
- когда входная переменная может принимать несколько значений (полихотомическая входная переменная);
- для случая непрерывной входной переменной.

Шансы и отношение шансов

Введем в рассмотрение понятие «шанс», который определяется как вероятность того, что событие произошло (шанс успеха), разделенная на вероятность того, что событие не произошло (шанс неуспеха). Шансы и вероятности содержат одну и ту же информацию, но по-разному выражают ее. Если вероятность того, что событие про изойдет, обозначить ρ, то шансы этого события будут равны $\rho/(1-\rho)$. Например, если вероятность выздоровления составляет 0,3, то шансы выздороветь равны $0{,}3/(1-0{,}3) = 0{,}43$. Если вероятность вытащить любую карту пиковой масти из колоды составляет 0,25, то шансы этого события равны $0{,}25/(1-0{,}25) = 0{,}33$.

Ранее мы обнаружили, что вероятность наличия заболевания у 72-летнего пациента составляет 61 %, соответственно, вероятность отсутствия заболевания 39 %. Таким образом, если обозначить шанс как O (Odds), то O = 0,61/0,39 = 1,56.

Мы также нашли, что вероятность наличия или отсутствия заболевания у 50-летнего пациента составляет 26 и 74 % соответственно. Тогда O = 0,26/0,74 == 0,35.

Заметим, что когда вероятность появления события выше, чем вероятность его отсутствия, то O > 1, а когда наоборот O < 1. Когда вероятности появления и отсутствия события равны, O = 1.

В бинарной логистической регрессии с дихотомической входной переменной вероятность того, что выходная переменная принимает значение y = 1 (событие произошло) при x = 1, может быть записано в виде [1]:

$$\frac{\rho(1)}{1-\rho(1)} = \frac{e^{\beta_0+\beta_1}/\left(1+e^{\beta_0+\beta_1}\right)}{1/(1+e^{\beta_0+\beta_1})} = e^{\beta_0+\beta_1}$$

Аналогично вероятность того, что выходная переменная принимает значение y = 1 (событие произошло) для наблюдений, в которых x = 0, будет [1]:

$$\frac{\rho(0)}{1-\rho(0)} = \frac{e^{\beta_0}/\left(1+e^{\beta_0}\right)}{1/(1+e^{\beta_0})} = e^{\beta_0}$$

Введем в рассмотрение так называемое отношение шансов, или отношение несогласия (odds ratio – OR), являющееся отношением шансов того, что событие произойдет, к шансам того, что событие не произойдет, В нашем случае это отношение шанса того, что выходная переменная примет значение 1 (событие произошло), к шансу того, что переменная примет значение 0 (событие не произошло),

То есть [1]:

$$OR = \frac{\rho(1)/(1-\rho(1))}{\rho(0)/(1-\rho(0))} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Данное отношение достаточно широко используется аналитиками, поскольку с его помощью хорошо выражается взаимосвязь между OR и коэффициентом b_1. Очевидно, что если отношение шансов равно 1, то есть шансы благоприятного и неблагоприятного исхода равны, то модель оказывается бесполезной, поскольку коэффициент $\beta_1 = 0$ и выход модели будут определяться только константой β_0. Таким образом, чем сильнее отношение шансов отличается от 1, тем более значимой будет модель. Если $OR < 1$, шансы благоприятного исхода меньше, чем шансы неблагоприятного (событие не произойдет), а если больше 1, то наоборот. Значения отношения шансов, близкие к 0, указывают на очень низкую вероятность благоприятного исхода.

Чтобы определить точность полученной оценки отношения шансов, используют стандартную ошибку отношения шансов, которая для случая дихотомической выходной переменной вычисляется с помощью выражения

$$E_{\text{ст}}(OR) = OR \times E_{\text{ст}}(b_1) = OR\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \qquad (1.11)$$

где $E_{\text{ст}}$ – стандартная ошибка оценивания соответствующего коэффициента регрессии, а значения $n_{11}, n_{12}, n_{21}, n_{22}$ – элементы четырехклеточной таблицы сопряженности признаков, отражающей все возможные состояния входной и выходной переменных,

Например, для задачи, в которой исследуется зависимость наличия некоторого заболевания от возраста пациента, таблица будет иметь следующий вид (табл. 8,17),

Таблица 1.10. Четырехклеточная таблица сопряженности признаков

Наличие заболевания, y	Возраст, x	
	$x < 50(0)$	$x \geq 50(1)$
Да (0)	$n_{11} = 21$	$n_{12} = 22$
Нет (1)	$n_{21} = 6$	$n_{22} = 51$

Тогда $E_{\text{ст}}(OR) = OR\sqrt{\frac{1}{21} + \frac{1}{22} + \frac{1}{6} + \frac{1}{51}} = 0{,}529 \times OR$

Границы доверительного интервала отношения шансов будут вычисляться по формулам [1]:

$$exp\left(b_0 \pm z \times E_{\text{ст}}(b_0)\right)$$
$$exp\left(b_1 \pm z \times E_{\text{ст}}(b_1)\right)$$

где z – критическое значение коэффициента Стьюдента, связанное с уровнем достоверности $(1-\alpha) \times 100\%$ (для $\alpha = 0{,}05 z \approx 1{,}96$);

$E_{\text{ст}}(b_0)$, $E_{\text{ст}}(b_1)$ – стандартные ошибки оценивания коэффициента регрессии на основе наблюдаемых данных. Если данный интервал не содержит $e^0 = 1$, то отношение шансов с достоверностью 95% является статистически значимым.

Ошибка $E_{\text{ст}}(b_1)$ для непрерывной переменной определяется как квадратный корень дисперсии оценки и вычисляется в процесс е оценки максимального правдоподобия. Ручные вычисления трудоемки и громоздки, поэтому значения ошибки берут из соответствующей компьютерной программы.

Часто отношение шансов используется для определения понятия относительно риска:

$$\text{относительный риск} = \frac{\rho(1)}{\rho(0)}$$

При низкой частоте событий значение отношения шансов приблизительно равно относительному риску.

Интерпретация модели для дихотомической переменной

Рассмотрим еще один пример. В сфере телекоммуникаций и предоставления услуг связи существует понятие «текучести» абонентской базы. Текучесть показывает, насколько часто клиенты отказываются от услуг данной компании и переходят к другим поставщикам услуг связи. Очевидно, что высокая текучесть проблема для компании. Поэтому представляют интерес исследования, цель которых заключается в выявлении причин ухода клиентов и в оценке вероятности ухода клиента с заданными показателями. На основе результатов таких исследований можно разработать методы работы с клиентами, позволяющие повысить их лояльность компании [3].

Фрагмент множества данных, собранных для такого исследования, представлен в табл. 1.11.

Таблица 1.11. Фрагмент набора данных «Текучесть абонентской базы»

Международные звонки	Голосовая почта	Количество голосовых сообщений	Использовано дневных минут	Количество звонков днем	Использовано вечерних минут	Количество звонков вечером	Использовано ночных минут	Количество ночных звонков	Междугородних разговоров, мин	Число междугородных звонков	Обращений в сервисную службу	Уход
Нет	Да	25	265.10	110	197,4	99	244,7	91	10,0	3	1	Нет
Нет	Да	26	161,60	123	195,5	103	254,4	103	13,7	3	1	Нет
Нет	Нет	0	243,40	114	121,2	110	162,6	104	12,2	5	0	Нет
Да	Нет	0	299,40	71	61,9	88	196,9	89	6,6	7	2	Нет
Да	Нет	0	166,70	113	148,3	122	186,9	121	10,1	3	3	Нет
Да	Нет	0	223,40	98	220,6	101	203,9	118	6,3	6	0	Нет
Нет	Да	24	218,20	88	348,5	108	212,6	118	7,5	7	3	Нет
Да	Нет	0	157,00	79	103,1	94	211,8	96	7,1	6	0	Нет
Нет	Нет	0	184,50	97	351,6	80	215,8	90	8,7	4	1	Нет
Да	Да	37	258,60	84	222,0	111	326,4	97	11,2	5	0	Нет
Нет	Нет	0	129,10	137	228,5	83	208,8	111	12,7	6	4	Да
Нет	Нет	0	187,70	127	163,4	148	196,0	94	9,1	5	0	Нет
Нет	Нет	0	128,80	96	104,9	71	141,1	128	11,2	2	1	Нет

Нет	Нет	0	156,60	88	247,6	75	192,3	115	12,3	5	3	Нет
…	…	…	…	…	…	…	…	…	…	…	…	…

Для описания признаков клиента, которые предположительно влияют на вероятность его ухода в другую компанию, используются следующие переменны.

- международные звонки – использует ли клиент международный тариф;
- голосовая почта – использует ли клиент службу отправки голосовых сообщений;
- количество голосовых сообщений – число отправленных голосовых сообщений;
- использовано минут днем – количество минут дневных разговоров, использованных клиентом;
- количество звонков днем – количество дневных звонков, сделанных клиентом;
- использовано вечерних минут – количество минут вечерних разговоров, использованных клиентом;
- количество звонков вечером – количество вечерних звонков, сделанных клиентом;
- использовано ночных минут – количество минут ночных разговоров, использованных клиентом;
- количество ночных звонков – количество ночных звонков, сделанных клиентом;
- минут международных разговоров – количество минут междугородних разговоров, использованных клиентом;
- число международных разговоров – число международных разговоров, сделанных клиентом;
- число обращений в сервисную службу – число обращений клиента в абонентскую сервисную службу.

Выходная переменная уход показывает, ушел ли клиент в другую компанию. Обозначим данную переменную как C.

Предположим, нужно оценить, как влияет использование или неиспользование клиентом службы голосовых сообщений (обозначим эту переменную VoiceMail) на вероятность его ухода к другому оператору связи. В табл. 1.12 указано, какое количество клиентов, использующих и не использующих голосовую почту, ушло и осталось,

Таблица 1.12. Сводная таблица ухода клиентов в зависимости от VoiceMail

	VoiceMail=Нет, x =0	VoiceMail=Нет, x =0	Всего
C = Нет, y = 0	2008	842	2850
C = Да, y = 1	403	80	483
Всего	2411	922	3333

Из таблицы видно, что из 2411 клиентов, не использующих голосовую почту 2008 клиентов остались, а 403 ушли в другую компанию. Из 922 клиентов, использующих голосовую почту, 80 клиентов ушли, а 842 остались. Оценим влияние использования голосовой почты на вероятность ухода клиента,

Дадим замечание. При использовании логистической регрессии сначала необходимо определить, какое событие и состояние выходной переменной связано с положительным исходом, а какое – с отрицательным. Например, если целью медицинского обследования является проверка наличия заболевания, то положительным исходом будет обнаружение болезни, а отрицательным ее отсутствие. Напротив, если обследование пациента про изводится после лечения, то целью является подтвердить результаты лечения. Тогда положительным исходом будет отсутствие заболевания, а отрицательным – ero наличие. В каждом случае положительный и отрицательный исход определяется логикой задачи. Для положительного исхода вероятность будет $\rho(x)$, а для отрицательного – $1-\rho(x)$. Что касается состояния бинарной выходной переменной, то 1 обычно связывают с положительным исходом, а 0 – с отрицательным [2].

Функция правдоподобия задается следующим образом:

$$l(\beta|x) = [\rho(0)]^{403} \times [(1-\rho(0))]^{2008} \times [\rho(1)]^{80} \times [(1-\rho(0))]^{842}$$

На основе данных, приведенных в табл. 1.12, вычислим шансы и отношения шансов. При этом, если клиент ушел, исход считается положительным, а если остался – отрицательным.

- Шанс ухода клиента, использующего голосовую почту: $O = \rho(1)/1-\rho(1) = 80/842 = 0{,}095/$

Шанс ухода клиента, не использующего голосовую почту: $O = 403/2008 = 0{,}201$

Тогда отношение шансов будет:

$$OR = \frac{\rho(1)/\big(1-\rho(1)\big)}{\rho(0)/\big(1-\rho(0)\big)} = \frac{80/842}{403/2008} = 0{,}473$$

Значения коэффициентов регрессии для модели с одной независимой переменной VoiceMail будут $b_0 = -1{,}606$ и $b_1 = -0{,}748$. Тогда $OR = exp(b_1) = exp(-0{,}748) = 0{,}47$, что совпадает с ранее вычисленным значением. Вероятность

ухода клиента, использующего (x = 1) или не использующего (x = 0) голосовую почту, будет равна [1]:

$$\hat{\rho}(x) = \frac{exp(\hat{g}(x))}{1 + exp(\hat{g}(x))}$$

а соответствующее логит-преобразование – $\hat{g}(x) = -1{,}606 - 0{,}748x$

Тогда для клиентов, использующих голосовую почту, вероятность ухода составит:

$$\hat{g}(x) = -1{,}606 - 0{,}748 \times 1 = -2{,}354$$

$$\hat{\rho}(x) = \frac{exp(-2{,}354)}{1 + exp(-2{,}354)} = 0{,}087$$

Иными словами, вероятность того, что клиент, использующий голосовую почту, откажется от услуг компании, составляет всего 8,7 %. Это меньше, чем общая доля клиентов, которые отказались от услуг компании, в исходном множестве данных, составившая 14,5 %. Следовательно, тот факт, что клиент использует голосовую почту, снижает вероятность его ухода.

Вероятность ухода также может быть вычислена непосредственно с помощью табл. 1.12:

$$P(C|VoiceMail) = 80/922 = 0{,}087$$

Для клиентов, не использующих голосовые сообщения, вероятность ухода оценивается как

$$\hat{g}(x) = -1{,}606 - 0{,}748 \times 0 = -1{,}606$$

$$\hat{\rho}(x) = \frac{exp(-1{,}606)}{1 + exp(-1{,}606)} = 0{,}167$$

Это ненамного выше 14,5%. Следовательно, если клиент не использует голосовую почту, вероятность его ухода несколько увеличивается.

Вероятность ухода также может быть вычислена непосредственно с помощью табл. 1.12:

$$P(C|VoiceMail) = 403/2411 = 0{,}167$$

Интерпретация модели для полихотомической входной переменной

Введем в рассмотрение переменную, которая указывает на количество обращений клиентов в абонентскую сервисную службу. Число обращений за поддержкой в сервисную службу зависит от числа проблем. А чем больше проблем, тем выше вероятность того, что клиент откажется от услуг компании. Чтобы использовать данную переменную, определим, какое число обращений в сервисную службу (обозначим эту переменную через CSC) можно рассматривать

как низкое, среднее и большое, Для этого введем фиктивные переменные (табл. 1.13).

Таблица 1.13. Квантование переменной CSC

	CSC_1	CSC_2
Низкое (0 или 1 вызов), CSC = Низкое	0	0
Среднее (2 или 3 вызова), CSC = Среднее	1	0
Высокое ($\geq$ 4 вызовов), CSC = Высокое	0	1

Таким образом, квантована переменная, указывающую на количество обращений клиентов в сервисную службу. При этом использовалась новая переменная CSC, которая порождает три фиктивные переменные. Выберем в качестве опорной категории CSC = Низкое.

В табл. 1.14 представлена зависимость значения выходной переменной от переменной CSC.

Таблица 1.14. Сводная таблица для переменной CSC

	CSC=Низкое	CSC=Среднее	CSC= Высокое	Всего
$C =$ Нет, $y = 0$	1664	1057	129	2850
$C =$ Да, $y = 1$	214	131	138	483
Всего	1878	1188	267	3333

Вычислим отношения шансов.

Для CSC = Среднее:

$$OR = \frac{138 \times 1664}{214 \times 1057} = 0{,}964$$

Для CSC = Высокое:

$$OR = \frac{131 \times 1664}{214 \times 129} = 7{,}90$$

Коэффициенты уравнения логистической регрессии будут равны $b_0 = -2{,}051$, $b_1 = -0{,}037$ и $b_2 = 2{,}118$.

Вероятность ухода клиента:

$$\hat{\rho}(x) = \frac{exp\big(\hat{g}(x)\big)}{1 + exp\big(\hat{g}(x)\big)}$$

с логит-преобразованием $\hat{g}(x) = -2{,}051 - 0{,}037CSC_1 + 2{,}118CSC_2$

Оценим вероятность ухода для клиентов с низким числом обращений в сервисный центр:

$$\hat{g}(x) = -2{,}051 - 0{,}0369891 \times 0 + 2{,}11844 \times 0 = -2{,}051$$

$$\hat{\rho}(x) = \frac{exp(-2{,}051)}{1 + exp(-2{,}051)} = 0{,}114$$

Таким образом, вероятность того, что клиент с низким числом обращений в сервисный центр откажется от услуг компании, составляет 11,4%. Это меньше, чем доля таких клиентов по всему исходному набору данных, которая составляет 14,5%. Следовательно, клиенты, которые редко обращаются в сервисный центр, менее склонны к уходу из компании.

Данная вероятность также может быть рассчитана непосредственно с помощью табл. 1.14:

$$P(C|CSC = \text{Низкое}) = 214/1878 = 0{,}114$$

Для клиентов со средним числом обращений:

$$\hat{g}(x) = -2{,}051 - 0{,}037 \times 1 + 2{,}118 \times 0 = -2{,}088$$

$$\hat{\rho}(x) = \frac{exp(-2{,}088)}{1 + exp(-2{,}088)} = 0{,}110$$

Оцененное значение показывает, что вероятность ухода клиентов, число обращений которых в сервисный центр мы определили, как среднее, примерно та же, что и для клиентов, редко обращающихся в сервисный центр.

Наконец, определим вероятность ухода клиентов, для которых зафиксировано большое число обращений в сервисный центр:

$$\hat{g}(x) = -2{,}051 - 0{,}037 \times 1 + 2{,}118 \times 1 = -0{,}003$$

$$\hat{\rho}(x) = \frac{exp(-0{,}003)}{1 + exp(-0{,}003)} = 0{,}50$$

Таким образом, вероятность ухода клиентов с большим числом обращений в сервисный центр составляет около 52%, а это более чем в три раза превышает вероятность ухода по выборке в целом. Очевидно, что компании необходимо выделять клиентов, которые 4 и более раза обращались в сервисный центр, и про водить с ними определенную работу.

Чтобы определить значимость переменной CSC = Среднее, выполним тест Вальда. Вычислим стандартную ошибку оценивания для данного коэффициента. Расчеты выполняются по той же схеме, что и для случая дихотомической входной переменной, но относительно опорной переменной. Это можно сделать, взяв данные непосредственно из табл. 1.14:

$$E_{\text{СТ}}\big(b(CSC = \text{Среднее})\big) = \sqrt{\frac{1}{1664} + \frac{1}{1057} + \frac{1}{214} + \frac{1}{131}} = 0{,}118$$

$$Z_W = \frac{b_1}{E_{\text{СТ}}(b_1)} = \frac{-0{,}037}{0{,}118} = -0{,}314$$

В данном случае ρ-значение $P(|z| \geq 0{,}314) = 0{,}753$ указывает на низкую значимость переменной. Поэтому в плане предсказания ухода клиента переменная

CSC = Среднее ненамного полезнее, чем CSC = Низкое. В то же время для CSC = Высокое имеем, что $b_1 = -2{,}118$, а $E_{\text{CT}}(b_1) = 0{,}1424$, поэтому:

$$Z_W = \frac{b_1}{E_{\text{CT}}(b_1)} = \frac{2{,}118}{0{,}1424} = 14{,}87$$

В этом случае ρ-значение $P(|z| \geq 14{,}87) \approx 0$. Очевидно, что переменная CSC = Высокое намного полезнее с точки зрения предсказания, чем переменная CSC = Низкое.

Доверительный интервал для отношения шансов между CSC = Высокое и CSC = Низкое вычисляется следующим образом:

$$CI = exp[2{,}118 \pm 1{,}96 \times 0{,}1424] = (e^{1{,}83}, e^{2{,}40}) = (6{,}23;\ 11{,}0)$$

Интервал не содержит значение $e^0 = 1$, поэтому отношение является статистически значимым.

Однако если рассмотреть доверительный интервал между CSC = Среднее и CSC = Низкое, получим:

$$CI = exp[0{,}037 \pm 1{,}96 \times 0{,}118] = (e^{-0{,}286}, e^{0{,}194}) = (0{,}77;\ 1{,}21)$$

Интервал содержит единицу, следовательно, отклонение отношения шансов от единицы является статистически незначимым на уровне $\alpha = 5\%$