2.6.2. Анализ выживаемости

Методы классификации, даже самые простые, такие как логистическая регрессия, — это мощный инструментарий для оценки вероятностей действий потребителей. Например, вероятность отклика на рекламное электронное письмо можно оценить, построив модель, использующую в качестве признаков атрибуты клиента, такие как количество покупок, и в качестве метки ответа — признак ответа клиента на предыдущее рекламное письмо. Этот подход широко используется на практике, но имеет несколько недостатков, которые мы подробно рассмотрим в последующих главах. Во-первых, во многих маркетинговых приложениях удобнее и эффективнее оценивать время до события, а не вероятность события. Например, для маркетинговой системы может быть полезнее оценить время до следующей покупки или время до отказа от подписки, чем вероятность этих событий. Во-вторых, маркетинговые данные очень часто включают записи с неизвестными или пропущенными результатами, которые нельзя должным образом учесть в моделях классификации. Возвращаясь к примеру с отказом от подписки, зачастую невозможно отличить клиентов, которые не сбежали, от клиентов, которые *пока* не сбежали, потому что мы строим предиктивную модель на определенный момент времени и не можем ждать бесконечно, пока появятся окончательные результаты для всех клиентов. Следовательно, мы знаем только результаты для клиентов, которые сбежали, и только их уверенно можем отметить как отрицательные образцы; остальные записи являются неполными, и нет никакой уверенности, что эти клиенты не сбегут в будущем, поэтому можно утверждать, что маркировка их как положительные или отрицательные образцы на самом деле не является действительной. Из этого следует, что было бы неправильно использовать модель классификации с бинарной переменной на выходе, определяемой на основе результатов, наблюдаемых в настоящий момент, и нам нужна другая статистическая структура для решения такого рода задачи.

Для медицинских и биологических исследований в свое время была разработана комплексная платформа для моделирования времени-до-события и обработки неполных данных. Основное внимание в исследованиях уделялось выживаемости людей после врачебной помощи, поэтому данную платформу назвали анализом выживаемости. Давайте опишем основные методы этой платформы, начав с базовой терминологии. Главная цель анализа выживаемости — оценить время до интересующего события и количественно объяснить, как это время зависит от параметров лечения, индивидуальных особенностей пациентов и других независимых переменных. В маркетинге аналогом лечения можно считать стимулирование или побуждение, например посредством рекламы. Роль события обычно играет покупка, активация, отказ от подписки или любое другое действие клиента, на которое маркетолог хотел бы повлиять. Отметим, что положительным результатом

лечения может быть приближение или отдаление момента наступления события, в зависимости от конечной цели. Рекламные объявления, например, направлены на стимулирование более ранних покупок, тогда как поощрительные предложения направлены на отдаление события отказа от подписки. В медицинских исследованиях, напротив, время обычно измеряется от диагноза до смерти, поэтому стандартная терминология анализа выживаемости предполагает, что событие соответствует некоторому отрицательному результату, что может сбивать с толку, когда имеет место обратное.

Как отмечалось выше, некоторые события могут быть неизвестны, в том смысле что результаты не наблюдались во время исследования. Эти пробелы в результатах могут иметь место из-за того, что результат не был известен на момент анализа (клиент еще не сделал покупку, но может сделать ее в будущем) или запись с информацией о клиенте была потеряна (например, из-за истечения срока действия cookie в веб-браузере). Записи с неизвестными результатами называют *цензурированными*. К моменту проведения анализа мы изначально имеем набор наблюдений, каждое из которых имеет время стимулирования и, не обязательно, время события.

Время между стимулированием и событием называют *временем выживаемости*. Мы можем преобразовать исходные наблюдения, упорядочив по времени стимулирования, чтобы наблюдения для k физических лиц (клиентов) представить в виде последовательности пар:

$$(t_1, \delta_1), ..., (t_k, \delta_k), \quad t_1 \leq ... \leq t_k, \tag{2.104}$$

где t обозначает временную метку события, а δ — индикатор, равный `1`, если наблюдение не цензурировано, и `0` в противном случае. Обычно предполагается непрерывность временной шкалы, но у двух клиентов может быть одинаковое время события, поэтому входные данные можно обобщить как

$$(t_1, d_1), ..., (t_n, d_n), \tag{2.105}$$

где n — число различных времен событий; d_i — общее число событий, наблюдаемых в момент времени t_i. Мы также предполагаем, что события не повторяются, то есть для каждого человека может произойти не более одного события. Для многих маркетинговых событий, таких как покупки, это предположение не является верным в буквальном смысле, но обычно мы можем обойти его, создавая отдельные модели для первого, второго и последующих событий, как мы обсудим в последующих главах. На данный момент нас интересует только распределение событий, и мы не пытаемся объяснить зависимость между временем выживаемости и параметрами стимулирования или особенностями клиентов.

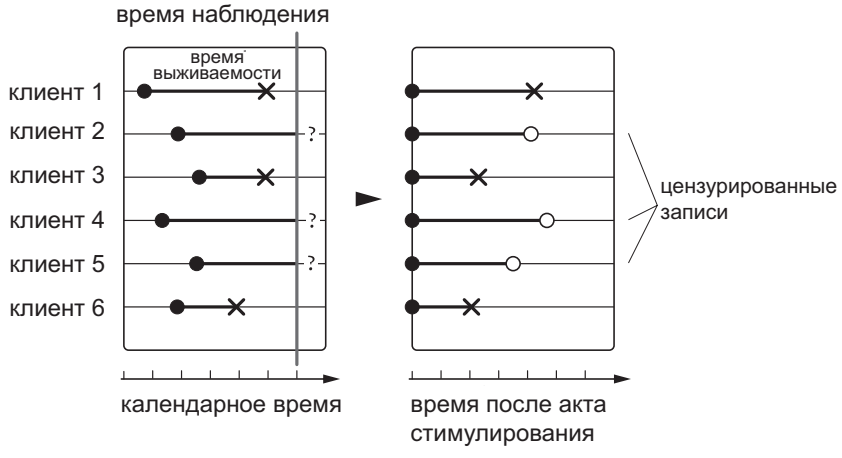

Рис. 2.12. Подготовка для анализа выживаемости. Закрашенные кружки соответствуют актам стимулирования. Крестики — соответствующим событиям. Незакрашенные кружки обозначают цензурированные записи

2.6.2.1. Функция выживаемости

Распределение времени выживаемости можно описать в терминах вероятности выживаемости, *функции выживаемости* $S(t)$, которая определяется как вероятность, что человек проживет от начального момента до момента времени t. Функция выживаемости — фундаментальная характеристика, описывающая динамику группы клиентов. Если функция выживаемости падает резко, событие с большой долей вероятности относительно быстро наступит для большинства клиентов. Если функция падает медленно, почти наверняка для большинства клиентов событие наступит в относительно отдаленной точке в будущем.

Обозначим время выживаемости клиента как T и его функцию плотности вероятности как $f(t)$. Кумулятивная функция распределения времени выживаемости, соответствующая вероятности события до момента времени t, будет тогда равна

$$F(t) = \Pr(T \le t) = \int_0^t f(\tau)\, d\tau, \tag{2.106}$$

а функцию выживаемости можно определить как

$$S(t) = \Pr(T > t) = 1 - F(t). \tag{2.107}$$

Значение функции выживаемости в момент времени t соответствует доле клиентов, для которых событие еще не наступило. Обратите внимание, что статистические

свойства времени выживаемости, такие как среднее, медиана и доверительные интервалы, можно оценить на основе кумулятивной функции распределения. Следовательно, эти свойства можно получить при наличии оценки функции выживаемости.

Функцию выживаемости можно вычислить по наблюдаемым данным с учетом как цензурированных, так и не цензурированных записей, исходя из предположения о том, что события независимы друг от друга. В этом случае кумулятивную вероятность выживаемости можно получить путем умножения вероятности выживаемости из одного интервала на вероятность в следующем. Более формально вероятность дожить до времени t можно оценить прямолинейно:

$$S_t = \frac{n_t - d_t}{n_t} = 1 - \frac{d_t}{n_t}, \tag{2.108}$$

где n_t — количество людей, для которых событие еще не наступило в момент времени t, а d_t — число людей, для которых событие уже наступило к моменту t. Путем перемножения вероятностей от начального момента времени до момента t можно оценить суммарную вероятность выжить, то есть функцию выживаемости:

$$\hat{S}(t) = \prod_{i \le t} \left(1 - \frac{d_i}{n_i} \right). \tag{2.109}$$

Эта оценка известна как оценка Каплана–Мейера (Kaplan–Meier), и можно доказать, что она является оценкой максимального правдоподобия [Kaplan and Meier, 1958]. Функция выживаемости равна 1 в нулевой момент времени, а затем, с увеличением времени, каждая точка данных вносит свой вклад в оценку. Проиллюстрируем оценку функции выживаемости на небольшом числовом примере.

ПРИМЕР 2.1

Предположим, что мы анализируем группу из 14 клиентов после того, как каждый из них получил рекламное письмо. Все письма были отправлены в разное время, и в ходе анализа фиксировалось время до первой покупки, прошедшее после отправки письма. Набор наблюдаемых данных выглядит следующим образом:

$$\begin{aligned} t &= \{2,\ 3,\ 3,\ 3,\ 4,\ 6,\ 7,\ 8,\ 12,\ 12,\ 14,\ 15,\ 20,\ 23\}, \\ \delta &= \{1,\ 1,\ 0,\ 1,\ 1,\ 1,\ 1,\ 0,\ 1,\ 1,\ 0,\ 1,\ 1,\ 1\}, \end{aligned} \tag{2.110}$$

где i-й элемент в множестве t — наблюдаемое время события для i-го клиента, измеренное в днях с момента отправки электронного письма. Множество δ содержит признак цензурированности наблюдения (`0`) или нецензурированности (`1`). Например, первый клиент совершил покупку на второй день после получения письма, а третий не совершил покупки к моменту анализа, хотя получил письмо за три дня до даты анализа. В этом контексте под вероятностью выживаемости подразумевается вероятность не совершить покупку к данному моменту времени. Многократно применяя формулу 2.109, получаем следующую последовательность:

$$\begin{aligned} &S(0)=1 \quad \text{(в начальный момент времени все клиенты «живы»)} \\ &S(2)=1-\frac{1}{14}=0{,}93 \\ &S(3)=S(2)\cdot\left(1-\frac{2}{13}\right)=0{,}79 \\ &\quad\ldots \end{aligned} \tag{2.111}$$

Этот результат соответствует ступенчатой *кривой выживаемости*, изображенной на рис. 2.13. Кривая выживаемости обобщает динамику группы клиентов, а кроме того, допустимо сравнивать кривые для различных групп. Например, можно построить кривые выживаемости для клиентов, которых стимулировали в рамках рекламной акции, и для тех, кому рекламные письма не рассылались, и графически оценить эффективность рекламной акции.

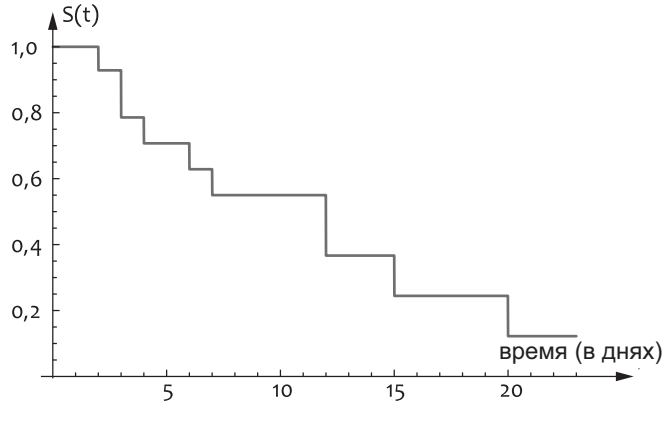

Рис. 2.13. Оценка функции выживаемости для набора данных в определении 2.110

2.6.2.2. Функция риска

Второе важное понятие в анализе выживаемости — *функция риска*. Если функция выживаемости фокусируется на вероятности, что событие не произойдет, то есть на *выживании*, то функция риска описывает риск наступления события. Как мы увидим позже, эта перспектива удобна для анализа влияния различных факторов, такие как параметры лечения, на время выживаемости.

Функция риска $h(t)$ определяется как мгновенная скорость риска, то есть вероятность события в бесконечно малом промежутке времени между t и $t+dt$ с учетом того, что человек дожил до времени t:

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t < T \le t + dt \mid T > t)}{dt}. \tag{2.112}$$

Функция риска имеет определенную связь с функцией выживаемости. Чтобы увидеть это, давайте сначала разложим условную вероятность в определении 2.112 на два фактора; отметьте, что один из них соответствует функции выживаемости:

$$\begin{aligned} h(t) &= \lim_{dt \to 0} \frac{\Pr(t < T \le t + dt)}{dt \cdot \Pr(T > t)} = \\ &= \lim_{dt \to 0} \frac{\Pr(t < T \le t + dt)}{dt \cdot S(t)} = \\ &= \lim_{dt \to 0} \frac{F(t + dt) - F(t)}{dt \cdot S(t)}. \end{aligned} \tag{2.113}$$

Затем вспомним, что функция плотности вероятности определяется как

$$f(t) = \lim_{dt \to 0} \frac{F(t + dt) - F(t)}{dt}. \tag{2.114}$$

Подставив это определение, а также определение функции выживаемости 2.107, мы получим следующий результат:

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = \\ &= -\frac{d}{dt} \log(1 - F(t)) = \\ &= -\frac{d}{dt} \log(S(t)). \end{aligned} \tag{2.115}$$

Решая это уравнение относительно $S(t)$, можно выразить функцию выживаемости как функцию $h(t)$:

$$S(t) = \exp\left(-H(t)\right), \tag{2.116}$$

где

$$H(t) = \int_0^t h(\tau)d\tau = -\log\left(S(t)\right) \tag{2.117}$$

называется кумулятивной функцией риска. Эта прямая связь позволяет переключаться между функциями риска и выживаемости в анализе.

2.6.2.3. Регрессионный анализ выживаемости

Базовые функции выживаемости и риска можно использовать для описания поведения группы клиентов или сравнения различных групп друг с другом. Но этого недостаточно для случаев, когда требуется понять и предсказать, как на выживаемость и риск влияют такие факторы, как маркетинговые действия и индивидуальные особенности клиента. Эта задача аналогична задачам классификации и регрессии, в том смысле что время выживаемости должно предсказываться как функция наблюдаемых факторов, то есть независимых переменных.

Предположим, что каждому индивидууму соответствует вектор $\boldsymbol{x}$, состоящий из p независимых переменных. То есть каждый индивидуум представлен тремя значениями:

- t — время выживаемости, или цензурированное время;
- δ — признак цензурирования, принимает значение 1 для наблюдаемых событий и 0 — для цензурированных;
- $\boldsymbol{x}$ — вектор признаков.

Набор входных данных содержит наблюдения для k клиентов:

$$(t_1, \delta_1, x_1), ..., (t_k, \delta_k, x_k), \quad t_1 \le ... \le t_k. \tag{2.118}$$

В маркетинговых приложениях вектор признаков может включать демографические и поведенческие свойства клиента, маркетинговые коммуникации с ним и т. д. Цель состоит в том, чтобы определить и обучить модель, которая выражает функции выживаемости и риска как функцию от $\boldsymbol{x}$. Поскольку $S(t)$ и $h(t)$ являются вероятностями, мы можем построить разные регрессионные модели выживаемо-

сти, предполагая разные распределения вероятностей и разные функциональные зависимости между признаками $\boldsymbol{x}$ и параметрами распределения.

Чаще всего в роли регрессионных моделей выживаемости используются модели *пропорциональных рисков*. Это семейство моделей основано на предположении, что единичное увеличение наблюдаемых факторов мультипликативно по отношению к степени риска, то есть

$$h(t \mid \boldsymbol{w}, \boldsymbol{x}) = h_0(t) \cdot r(\boldsymbol{w}, \boldsymbol{x}), \tag{2.119}$$

где $h_0(t)$ является базовым риском, r — отношение рисков, увеличивающее или уменьшающее базовый риск в зависимости от факторов, а $\boldsymbol{w}$ — вектор параметров модели. Обратите внимание, что базовый риск зависит не от конкретного человека, а от отношения рисков. Другими словами, отношение рисков определяет, как свойства индивидуума, закодированные в векторе признаков, влияют на уровень риска. Отношение рисков не может быть отрицательным, так как коэффициент риска не отрицателен, поэтому он обычно моделируется как экспоненциальная функция:

$$h(t \mid \boldsymbol{w}, \boldsymbol{x}) = h_0(t) \cdot \exp(\boldsymbol{w}^T \boldsymbol{x}). \tag{2.120}$$

Эту модель можно считать линейной по отношению к логарифму отношения рисков индивидуума к базовому риску:

$$\log r(\boldsymbol{w}, \boldsymbol{x}) = \log \frac{h(t \mid \boldsymbol{x})}{h_0(t)} = \boldsymbol{w}^T \boldsymbol{x}. \tag{2.121}$$

Что касается базового риска $h_0(t)$, у нас на выбор есть два варианта: параметрический и непараметрический. Параметрический подход предполагает, что риск подчиняется определенному распределению. В этом случае мы получаем полностью параметрическую модель, которую необходимо обучить на исходных данных путем поиска оптимальных значений параметров $\boldsymbol{w}$ и параметров распределения. Недостатком этого подхода является предположение, что базовый риск изменяется во времени определенным образом, поэтому мы должны быть уверены, что выбранное распределение соответствует данным. С другой стороны, параметрический подход сглаживает зашумленные данные и обеспечивает простую модель базового риска.

Второй вариант заключается в использовании непараметрической модели базового риска, которую можно получить на основе данных с использованием оценки Каплана–Мейера или других методов. Это приводит к полупараметрической модели для общего риска, где параметрическая часть определяется выражением 2.120, а базовый риск $h_0(t)$ представляет непараметрическую часть. Это решение известно

как модель пропорциональных рисков Кокса [Cox, 1972]. Преимущество модели Кокса, как мы увидим ниже, заключается в возможности оценить коэффициенты риска без необходимости вычислять базовую функцию риска или делать какие-либо предположения о структуре базового риска. Это делает ее очень удобной для применений, где требуется определить только факторы риска, а не абсолютные значения. Недостатком модели Кокса является необходимость определения базового риска с помощью параметрических методов. Важно также иметь в виду, что модель Кокса относится к семейству моделей пропорциональных рисков и, следовательно, основана на предположении о пропорциональности рисков, что может быть верно или неверно для наблюдаемых данных. Модель Кокса широко используется во многих областях, включая маркетинг, и мы будем использовать ее в качестве основного инструмента для анализа выживаемости в следующей главе.

Наш следующий шаг — определить параметры модели Кокса по данным. Стандартное решение этой задачи состоит в том, чтобы получить правдоподобие модели, а затем найти параметры, максимизирующие его. Проблема, однако, в том, что наблюдения могут оказаться цензурированными, а это требует от нас указать, как такие записи должны учитываться в определении правдоподобия. Прежде всего отметим, что каждое наблюдение вносит свой вклад в величину подобия. Если i-е наблюдение цензурировано, это означает вероятность дожить до t_i:

$$L_i(\boldsymbol{w}) = S(t_i \mid \boldsymbol{w}, \boldsymbol{x}). \tag{2.122}$$

Если наблюдение не цензурировано, оно вносит вклад в вероятность возникновения события в момент t_i, которая определяется с помощью функции плотности вероятности времени выживаемости:

$$L_i(\boldsymbol{w}) = f(t_i) = h(t_i \mid \boldsymbol{w}, \boldsymbol{x}) S(t_i \mid \boldsymbol{w}, \boldsymbol{x}). \tag{2.123}$$

То есть полное правдоподобие можно выразить так:

$$L_i(\boldsymbol{w}) = \prod_{i=1}^{k} h(t_i \mid \boldsymbol{w}, \boldsymbol{x})^{\delta_i} S(t_i \mid \boldsymbol{w}, \boldsymbol{x}). \tag{2.124}$$

Мы не сможем максимизировать это выражение с использованием численных методов, не указав форму базового риска. Однако можно аппроксимировать полное правдоподобие с помощью другой меры, называемой частичным правдоподобием. Для начала введем понятие *группы риска* в момент времени t, которое определяется как множество лиц с риском наступления события в момент t, то есть лиц, для которых событие еще не наступило:

$$R(t) = \{i : t_1 \geq t\}. \tag{2.125}$$