

# Модели кредитного и поведенческого скоринга

## 1 Введение

Назначение кредитного скоринга — автоматизированное принятие решений по выдаче кредитов частным лицам. Слово «скоринг» происходит от английского *score*: во многих используемых системах на основании данных, указанных в заявке клиента, вычисляется специальное число (счёт), которое и используется затем в качестве критерия кредитоспособности заёмщика. Подробные обзоры методов скоринга содержатся в статьях [13, 25, 15].

В зависимости от типа используемых входных данных о потенциальном заёмщике выделяют следующие типы скоринга.

- *Кредитный скоринг* (скоринг по данным заявки, *application scoring*) — принятие решения о выдаче кредита новым клиентам по данным, указанным в заявке.
- *Поведенческий скоринг* (*behaviour scoring*) — динамическая оценка состояния кредитоспособности существующего заёмщика, основанная на данных об истории транзакций по его счетам (график погашения задолженности, оборот по текущим счетам, запрос новых кредитов и т.п.). По результатам оценки может определяться текущий лимит кредита для заёмщика; меры, принимаемые в случае задержки платежей; маркетинговые ходы, которые могут быть направлены на клиента.

При построении модели скоринга также могут использоваться данные различных видов, и здесь можно провести следующее разделение.

- *Априорный скоринг* — построение модели по статистическим данным (макроэкономические показатели, государственная статистика, результаты переписи и т.п.). Эти данные используются для оценки параметров модели заёмщика, которая в свою очередь применяется для определения его кредитоспособности.
- *Апостериорный скоринг* — построение модели по историческим данным о клиентах кредитной организации. Такие данные обычно представляют собой таблицу, в которой для каждого из клиентов, получившего кредит, указаны данные его заявки, а также исход — погашение долга, досрочное погашение или дефолт. Таким образом, производится построение модели, по возможности лучше предсказывающей исход на основании анкетных данных.

Система скоринга может использоваться не только на стадии продажи кредитного продукта, но и при его проектировании, поскольку с её помощью можно определить проанализировать кредитоспособность группы потенциальных заёмщиков, под которую проектируется продукт, и, выделив основные качества заёмщиков, способствующие снижению риска, направить основные маркетинговые усилия именно на таких заёмщиков.

## 2 Кредитный скоринг

Кредитный скоринг — это метод разграничения различных групп потенциальных клиентов в условиях, когда доступна информация не о параметрах, разделяющих эти группы, а только о некоторых вторичных переменных. Идея разделения групп по вторичным признакам высказал в 1936 г. R. A. Fisher [10] (он предложил метод определения разновидностей ириса по результатам измерений размеров частей растений), а несколькими годами позже D. Durand [9] предложил использовать тот же самый подход для распознавания заёмщиков с высокой и низкой вероятностью дефолта.

Доступная информация о потенциальных заёмщиках содержится в заполненных ими анкетах. Такие факторы, как годовой доход, размер непогашенного долга, владение недвижимостью или автомобилем, стаж работы на последнем месте, возраст и т.п. потенциально связаны с кредитоспособностью и поэтому могут оказаться входными переменными в скоринговой модели. Однако вследствие корреляций между этими факторами, в итоговой модели окажутся не все из них, а только некоторые. Согласно Fair & Isaac, в ходе построения модели могут участвовать 50–60 переменных, тогда как в конечную модель попадут не более 8–12 из них [13].

### 2.1 Модели кредитного скоринга

В настоящее время для кредитного скоринга используются методы статистики (дискриминантный анализ, линейная регрессия, логистическая регрессия, деревья классификации), исследования операций (линейное программирование, нелинейная оптимизация) и искусственного интеллекта (нейронные сети, экспертные системы, генетические алгоритмы, методы ближайших соседей, байесовские сети, логико-вероятностные методы). Указанные методы могут применяться как по отдельности, так и в различных комбинациях. Обилие совершенно разных методов для решения одной и той же задачи объясняется чисто прагматическим подходом: использовать то, что работает, а не пытаться объяснить причину дефолтов или зависимость от макроэкономических показателей.

Для построения модели берётся выборка данных по существующим заёмщикам, обычно не менее нескольких тысяч записей. По каждому заёмщику необходимы анкетные данные и кредитная история за определённый период, как правило от одного года до двух лет. Относительно каждой кредитной истории решается, является ли она «хорошей» или «плохой»; обычно признаком «плохой» кредитной истории является факт задержки платежей на три и более месяцев. По части клиентов такое решение принять трудно (скажем, по тем, кто пропустил платежи за два месяца). Таких клиентов помещают в группу «средние». В зависимости от используемого метода скоринга, при построении модели могут все три указанные группы, либо «средние» клиенты просто удаляются из выборки.

На качество модели сильно влияет выбор периода, за который рассматривается кредитная история (т.е. промежуток времени между заполнением анкеты и классификацией клиентов на «плохих» и «хороших»). Считается, что вероятность дефолта

как функция времени от прихода в кредитную организацию вначале сильно колеблется, и только после года начинает стабилизироваться, поэтому использование меньшего периода приведёт к недооцениванию вероятности дефолта. Напротив, использование периода более двух лет нежелательно в силу того, что за это время могут произойти определённые социально-экономические изменения, так что вновь приходящие клиенты будут слишком сильно отличаться по своим характеристикам от тех, на которых была построена модель.

Соотношение числа «плохих» и «хороших» клиентов в обучающей выборке также влияет на качество системы скоринга. Для одних методов (например, линейная регрессия) предпочтительно использование равных долей «плохих» и «хороших», тогда как другие методы (деревья классификации, байесовские сети) требуют, чтобы выборка отражала реальное соотношение «плохих» и «хороших» клиентов.

Проблему кредитного скоринга можно рассматривать как задачу классификации: зная ответы на вопросы анкеты  $x \in A$ , определить, к какой группе относится заёмщик:  $x \in A_G$  для «хороших» клиентов, и  $x \in A_B$  для плохих. При этом необходимо понимать, что абсолютно точная классификация принципиально невозможна. Хотя бы потому, что один и тот же набор ответов может быть дан как «хорошим», так и «плохим» клиентом (напомним, что вопросы в анкете представляют собой вторичные характеристики заёмщиков). Хотелось бы, однако, построить такую модель, которая производила бы правильную классификацию как можно в большем числе случаев.

Большинство статистических методов приводят к построению правила классификации, основанного на линейной скоринговой функции. Оно может быть получено использованием различных подходов.

**Байесовский подход.** Предположим, что известны следующие функции и величины:

- $p_G$  — доля «хороших» клиентов, и, соответственно,  $p_B = 1 - p_G$  — доля «плохих» клиентов;
- $p(x|G)$ ,  $p(x|B)$  — вероятности того, что соответственно «хороший» и «плохой» клиент дадут ответы  $x$ ;
- $L$  — потери от того, что «хороший» клиент будет классифицирован как «плохой»;
- $D$  — потери в случае дефолта, т.е. от классификации «плохого» клиента как «хорошего».

По формуле полной вероятности можно тогда вычислить  $p(x) = p(x|G)p_G + p(x|B)p_B$  — вероятность того, что клиент даст ответы  $x$ , и затем по теореме Байеса определить

$$q(G|x) = \frac{p(x|G)p_G}{p(x)}, \quad q(B|x) = \frac{p(x|B)p_B}{p(x)},$$

то есть вероятности того, что клиент с данными ответами будет «хорошим» или «плохим».

Средние потери в расчёте на одного клиента составляют

$$EL = L \sum_{x \in A_B} q(G|x)p(x) + D \sum_{x \in A_G} q(B|x)p(x) = L \sum_{x \in A_B} p(x|G)p_G + D \sum_{x \in A_G} p(x|B)p_B,$$

и являются минимальными при выборе

$$A_G = \{x \mid Dp(x|B)p_B \leq Lp(x|G)p_G\}.$$

Если далее предположить нормальность распределений  $p(x|G)$ ,  $p(x|B)$  с общей ковариационной матрицей, то получается линейное правило

$$A_G = \{x \mid w_1x_1 + \dots + w_nx_n > c\}.$$

Такое правило фактически означает, что каждому набору ответов  $x$  присваивается «счёт»  $s(x)$  по формуле

$$s(x) = w_1x_1 + \dots + w_nx_n,$$

и означает, что в  $s(x)$  содержится достаточно информации для различения класса клиента. Таким образом, размерность задачи уменьшается с  $n$ , с распределениями  $p(x|G)$ ,  $p(x|B)$ , до 1 с распределениями  $p(s|G)$ ,  $p(s|B)$ . Нахождение правила классификации сводится к поиску оптимального порога  $c$  из задачи оптимизации

$$L \sum_{s < c} p(s|G)p_G + D \sum_{s \geq c} p(s|B)p_B \rightarrow \min.$$

**Множественная регрессия.** Другой способ получить линейную скоринговую функцию — использовать одну из разновидностей линейной регрессии. В простейшем варианте определяется зависимая переменная  $Y$ , принимающая значения 1 в случае «хорошего» клиента и 0 в случае «плохого», и затем используется метод наименьших квадратов для нахождения весов  $w_i$ , обеспечивающих наилучшее предсказание значения  $Y$ .

При использовании линейной регрессии фактически делается попытка связать вероятность дефолта  $p$  со значениями ответов на вопросы линейной функцией:

$$p = w_0 + w_1X_1 + \dots + w_nX_n.$$

Здесь есть очевидное несоответствие: левая часть представляет собой вероятность и должна изменяться от 0 до 1, тогда как правая может принимать любые значения. Для преодоления этой трудности само значение вероятности заменяется некоторой функцией от него, так чтобы эта функция была монотонной принимала все вещественные значения.

Логистическая регрессия заменяет вероятность дефолта на логарифм шансов дефолта:

$$\log \frac{p}{1-p} = w_0 + w_1 X_1 + \dots + w_n X_n = s(X).$$

Таким образом, счёт связан здесь не с вероятностью дефолта  $p(B|x)$ , а с отношением  $p(B|x)/p(G/x)$ :

$$s(x) = \log \frac{p(B|x)}{p(G|x)}.$$

Другим примером является пробит-анализ:

$$\Phi^{-1}(p) = w_0 + w_1 X_1 + \dots + w_n X_n,$$

где  $\Phi(x)$  — функция стандартного нормального распределения.

**Дискриминантный анализ.** Используя линейные скоринговые функции, можно оценить вероятность принадлежности клиента тому или иному классу, определив

$$\begin{aligned} s_1(x) &= w_0^1 + w_1^1 X_1 + \dots + w_n^1 X_n, \\ s_2(x) &= w_0^2 + w_1^2 X_1 + \dots + w_n^2 X_n, \\ &\dots \\ s_m(x) &= w_0^m + w_1^m X_1 + \dots + w_n^m X_n. \end{aligned}$$

Далее необходимо просто выбрать тот класс, которому соответствует больший счёт. В случае двух классов это даёт тот же самый результат, что и линейная регрессия.

**Линейное программирование.** Ещё один способ получить линейную скоринговую функцию заключается в использовании линейного программирования. Пусть есть набор данных, включающий в себя ответы на вопросы  $x_{i1}, \dots, x_{in}$  для каждого из  $N$  клиентов и индикаторы наличия дефолта  $Y_i$ . Идеально было бы найти линейную функцию  $s(x)$ , которая бы полностью разделяла «хороших» и «плохих» клиентов. Однако поскольку это обычно невозможно, будем искать  $s(x)$  из соображений минимизации ошибки  $a_i$  в каждом конкретном случае:

$$\begin{cases} a_1 + \dots + a_N \rightarrow \min, \\ w_1 x_{i1} + \dots + w_n x_{in} \geq c - a_i, & Y_i = 0, \\ w_1 x_{i1} + \dots + w_n x_{in} \leq c + a_i, & Y_i = 1, \\ a_i \geq 0. \end{cases}$$

Здесь минимизация производится по переменным  $(w_1, \dots, w_n, c, a_1, \dots, a_N)$ . Можно также минимизировать максимальную возможную ошибку, заменив все  $a_i$  на одну переменную  $a$ .

Преимущество метода линейного программирования состоит в возможности включить в программу дополнительные ограничения. Например, если кредитный продукт направлен на людей молодого возраста, можно наложить дополнительное ограничение на соответствующий вес  $w_i \geq w_i^0$ , так что скоринговая функция будет отдавать предпочтение людям этого возраста.

**Генетические алгоритмы** представляют собой метод оптимизации, не накладывающий стандартных ограничений на целевую функцию (гладкость, выпуклость и т.п.). В применении к задаче скоринга генетический алгоритм состоит в генерации начального множества скоринговых функций, с последующим применением к функциям из этого набора операций «скрещивания» и «мутации», а также отбрасыванием наименее пригодных функций.

**Деревья классификации.** В отличие от предыдущих методов, классификационные деревья (иначе — рекурсивные алгоритмы разбиения) не предназначены для построения линейной скоринговой функции. Вместо этого они последовательно разбивают клиентов на группы по одной из переменных, так чтобы эти группы насколько возможно отличались по величине кредитного риска. Процесс разбиения продолжается до тех пор, пока оставшиеся группы не становятся настолько малы, что следующее разбиение не приведёт к статистически значимому различию в уровне риска. Каждому листу дерева затем приписывается определённая категория клиентов.

На рис. 1 изображён пример дерева классификации<sup>1</sup>.

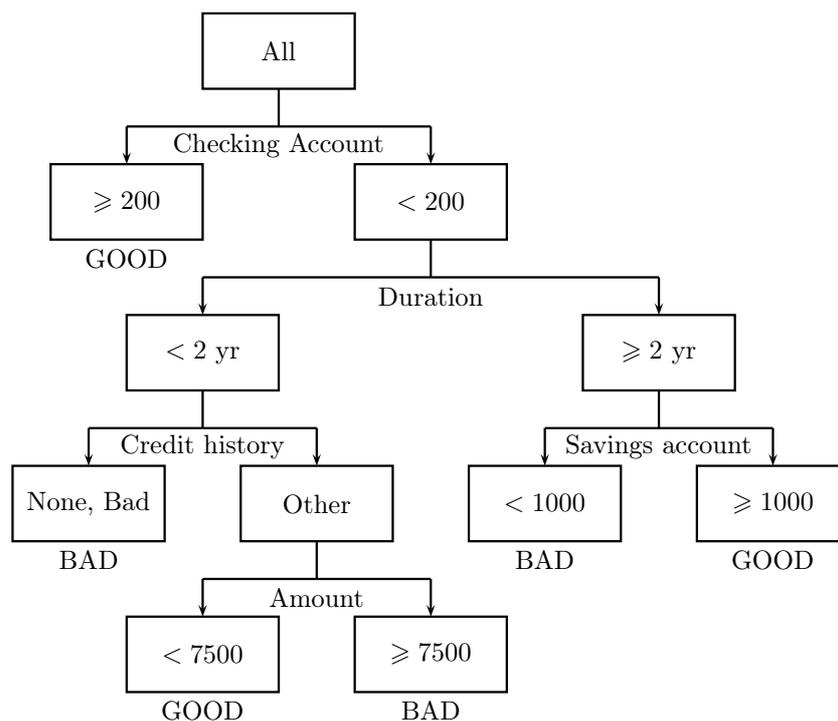


Рис. 1: Дерево классификации

Модификацией деревьев классификации являются *деревья моделей*, где каждому

<sup>1</sup>Здесь и далее в примерах использован набор данных German из UCI Machine Learning Repository [2].

листу соответствует не определённая категория, а своя линейная регрессионная модель. Это позволяет сначала достаточно грубо разделить клиентов на основные группы, а потом для каждой группы воспользоваться регрессионной моделью (см. пример в [17]).

**Метод ближайших соседей** является примером подхода «ленивого обучения», когда обучение фактически сводится к добавлению нового случая в базу данных. На пространстве анкетных данных выбирается некоторая метрика, которая затем используется для нахождения ближайших исторических анкетных данных к данным нового клиента. При этом клиент классифицируется в зависимости от того, к какому классу принадлежат большинство его соседей по этой метрике.

Преимуществом метода ближайших соседей является исключительная простота обучения модели: при поступлении новых данных их необходимо просто добавить к существующим, а по прошествии времени можно удалять старые данные, чтобы учитывать изменения, происходящие с течением времени.

К недостаткам следует отнести необходимость полного просмотра базы данных при классификации нового клиента (что может занимать довольно много времени в случае больших объёмов информации, скажем, более  $10^5$ – $10^6$  записей). Кроме того, выбор хорошей метрики также является нетривиальной задачей, и если производить его на основе исторических данных, то теряется основное привлекательное свойство алгоритма — отсутствие необходимости предварительного анализа данных и построения модели.

**Нейронные сети** могут рассматриваться как метод нелинейной регрессии и показывают очень хорошие результаты в задаче кредитного скоринга. Они чаще применяются для скоринга компаний, когда имеется меньше данных, чем для скоринга частных лиц.

Недостатком нейронных сетей является то, что веса связей, определённые в результате обучения, не имеют никакой интерпретации в терминах кредитного риска. Следовательно, практически невозможно объяснить предсказание, полученное с помощью нейронной сети, а также провести анализ чувствительности, чтобы выделить наиболее значимые параметры.

**Байесовские сети** позволяют представлять многомерные распределения  $p(x|G)$ ,  $p(x|B)$  в виде комбинации нескольких распределений более низкой размерности [21]. При этом в ходе построения модели могут быть учтены причинно-следственные связи (благодаря чему они широко распространены в таких областях как медицинская диагностика, поиск технических неисправностей и т.п.). Хотя в кредитном скоринге и других задачах классификации как правило рассматриваются переменные, которые не связаны друг с другом причинно, однако можно предположить существование скрытых переменных, обуславливающих ту или иную условную зависимость или независимость. Более подробно этот подход описан в п. 2.6.

Безусловным преимуществом байесовских сетей является возможность вывода по неполным данным. Если информация о потенциальном заёмщике не является полной, то алгоритм вывода по байесовской сети вычислит вероятность дефолта, основываясь лишь на доступных данных (что эквивалентно усреднению по переменным, чьи значения неизвестны).

Байесовский подход позволяет также относительно легко производить интеграцию системы апостериорного скоринга с другими используемыми моделями, в частности, с системой априорного скоринга, чтобы учесть изменения в распределении дефолтов, связанные с изменением экономической ситуации.

**Логико-вероятностный подход** [1] основан на сочетании методов математической логики и теории вероятностей. Этот подход можно рассматривать как специальный вид нелинейной регрессии. В качестве параметров модели здесь выступают вероятности наступления дефолта, связанного с тем или иным ответом на вопрос анкеты. Оцениваемая величина — вероятность дефолта. Для идентификации значений параметров используется специально адаптированные алгоритмы оптимизации (основной трудностью является то, что целевая функция — количество правильно классифицированных случаев — является целочисленной, поэтому не работают стандартные алгоритмы оптимизации типа наискорейшего спуска).

## 2.2 Информация на выходе системы скоринга

В зависимости от используемой модели система скоринга может выдавать на выходе следующие данные.

- *Класс клиента.* В самом простом случае — «хороший» или «плохой»; первым кредит выдается, вторым — нет. В более сложных случаях может быть ещё несколько промежуточных классов — кредит выдается, но на других условиях (большой процент, меньший срок, или дополнительные гарантии), либо кредитоспособность оценивается вручную, с использованием дополнительной информации.
- *Апостериорное распределение класса клиента.* Для каждого класса указывается вероятность, с которой данный клиент принадлежит этому классу. Можно выбрать класс с наибольшей вероятностью, либо усреднить какие-либо показатели по каждому классу. Данный вариант предпочтительнее первого, поскольку в распределении содержится существенно больше информации. Например, если классов всего два, то в первом варианте клиент будет считаться «хорошим» как при распределении 90%/10%, так и при 55%/45%. Очевидно, однако, что во втором случае степень уверенности в кредитоспособности существенно ниже, чем в первом.
- *Вероятность дефолта.* Для принятия решения остаётся только сравнить эту вероятность с пороговым значением допустимой вероятности дефолта. Последняя

определяется так, чтобы с учетом этой вероятности и процентной ставки банк в среднем не терпел убытков в случае выдачи кредита.

- *Счёт (score)* — количественная оценка кредитоспособности потенциального заёмщика (чем больше счёт, тем последняя выше). Обычно счёт пропорционален вероятности или шансам успешного возврата кредита, поэтому либо по счёту определяется вероятность дефолта и на этом основании принимается решение, либо, в зависимости от того, в какой промежуток попал счёт, определяется класс клиента и на основании этого — условия выдачи кредита.

Если скоринговая система так или иначе определяет вероятность дефолта заёмщика  $p_{\text{деф}}$ , то необходимо установить пороговую (допустимую) вероятность дефолта. Это можно сделать, вычислив средние потери от дефолта, и введя ограничение, что эти потери не могут быть больше дохода в случае успешного возврата кредита.

Потери в случае выдачи кредита неплатежеспособному заёмщику, потери в случае отказа «хорошему» клиенту образуют так называемую *матрицу штрафов* (misclassification costs matrix). В этой матрице столбцам соответствуют фактические (реальные) классы клиентов, а строкам — предсказанные. На диагонали находятся нули (т.е. за правильно предсказанный класс штраф отсутствует), а во всех остальных ячейках указывается штраф за предсказание одного класса (строка), если на самом деле клиент принадлежит другому классу (столбец). Например, в случае матрицы

	good	bad
good	0	5
bad	1	0

штраф составляет 1 единицу за отказ «хорошему» клиенту и 5 единиц за выдачу кредита «плохому» клиенту. Как видно, в случае двух классов матрица штрафов содержит два значения — штрафы за ошибки первого и второго рода.

Элементы матрицы могут быть определены исходя из порогового значения вероятности дефолта  $p^*$  (выше которого кредиты выдаваться не должны). Примем, что штраф за отказ хорошему клиенту равен 1, а  $n$  — штраф за приём плохого клиента. Тогда

$$p^* = \frac{1}{1+n}, \quad n = \frac{1-p^*}{p^*}.$$

Например,  $n = 19$  при  $p = 5\%$  и  $n = 9$  при  $p = 10\%$ . Для матрицы, приведённой выше,  $p^* = \frac{1}{6} \approx 17\%$ .

Матрица штрафов используется в ходе оценки качества используемых моделей, а также является частью входных данных некоторых алгоритмов (например, при построении деревьев классификации).

## 2.3 Выбор и оценка скоринговых моделей

Кратко опишем общую методологию оценки моделей добычи данных, следуя работам Y. Liu [16, 17].

**Выбор модели.** Процесс построения скоринговой системы можно условно разбить на три этапа:

1. *Формулировка задачи и подготовка данных.* С помощью экспертов в конкретной области формулируется задача скоринга, производится сбор и предварительная обработка данных.
2. *Анализ данных и построение модели.* Производится поиск оптимальной модели для решения поставленной задачи. Необходимо оценить точность работы различных моделей и выбрать наилучшую из них.
3. *Применение и валидация модели.* Модель применяется для реального принятия решений, при этом производится оценка её точности на фактических данных. По прошествии времени модель должна перестраиваться, чтобы отражать произошедшие изменения.

Несмотря на то, что кредитный скоринг предназначен для автоматического принятия решения по выдаче кредитов, сам процесс построения модели для скоринга не может обходиться без непосредственного участия человека на каждом из этапов.

**Точность классификации.** Наиболее очевидным критерием точности является процент неверной классификации, или интенсивность ошибок:

$$\text{интенсивность ошибок} = \frac{\text{число случаев неверной классификации}}{\text{общее число случаев}}.$$

Для задачи классификации с двумя классами это число должно быть между нулём (все случаи классифицированы корректно) и интенсивностью ошибок классификации по умолчанию (присваивающей во всех случаях класс, которому принадлежит большинство клиентов). По ряду причин построенная модель должна иметь меньшую интенсивность ошибок, чем классификация по умолчанию, при этом в реальных приложениях не существует моделей с нулевой интенсивностью ошибок.

*Реальная интенсивность ошибок* (true error rate) определяется тестированием модели на настоящих данных. Она не может быть определена до тех пор, пока модель не будет протестирована на большом количестве реальных случаев. Следовательно, в ходе построения модели этот показатель необходимо как-либо оценить.

*Собственная интенсивность ошибок* (apparent error rate) определяется как интенсивность ошибок на наборе данных, который был использован для обучения модели. Однако она не является надёжной оценкой реальной интенсивности, поскольку низкое её значение может означать, что модель является просто-напросто «подгонкой» (overfitting) результата классификации под данные в обучающем наборе (например, у метода ближайших соседей всегда будет нулевая собственная интенсивность ошибок). В этом случае можно ожидать весьма посредственных результатов при применении модели к реальным данным.

Для оценки собственной интенсивности ошибок применяется метод «удержания» тестовых данных: исходный набор данных разделяется на «обучающий» (используемый для построения модели) и «тестовый» (используемый для оценки точности) наборы. Предполагается, что тестовый набор выделяется случайным образом, независимо от самих данных. Определённая таким образом интенсивность ошибок называется *тестовой интенсивностью ошибок*. Обычно величина тестового набора составляет около 30% от всех данных. При величине тестового набора в 1000 записей тестовая интенсивность ошибок уже является статистически точной оценкой реальной интенсивности.

Представляет также интерес точность классификации при условии изменений, происходящих в населении с течением времени. Для этого имеющиеся данные упорядочиваются по дате заполнения анкеты, и затем в качестве обучающего набора используется первая часть списка, а в качестве тестового — вторая. Это также отчасти решает проблему излишней «подгонки» под обучающие данные.

Отметим, что сравнение результатов точности классификации различными методами показывает, что они практически совпадают [25]. Это можно объяснить эффектом «плоского максимума» [18]: существенные изменения весов в окрестности оптимальной скоринговой функции приводят к незначительным отклонениями в точности прогнозов.

**Критерии применимости.** Хотя точность классификации и является важным критерием выбора скоринговой модели, необходимо также принимать во внимание ряд других качеств.

- *Скорость работы.* Необходимо оценить время, требуемое для обучения и для принятия решения в соответствие с моделью. Приведём два крайних примера. С одной стороны, метод ближайших соседей исключительно быстр в обучении (просто добавляется ещё одна запись), но для принятия решения требуется полный перебор случаев в базе данных, что может занять много времени. Напротив, нейронные сети требуют минимальных вычислений для классификации одного случая, но при этом их обучение является *NP*-трудной задачей, поэтому алгоритмы обучения требуют экспоненциального времени.
- *Прозрачность и интерпретируемость.* Прозрачность моделей становится важной, когда модель необходимо объяснить кредитным аналитикам. Часто счёт, выдаваемый системой, используется как один из критериев принятия решения квалифицированным кредитным офицером, поэтому модель должна быть в достаточной мере понятной. Наиболее прозрачными моделями являются, по всей видимости, основанные на линейной скоринговой функции. Напротив, нейронные сети действуют как «чёрный ящик» и не предоставляют никаких объяснений результатов классификации, что препятствует использованию таких систем на практике, когда кредиторам требуется объяснять, почему они не выдали тот или иной кредит.

- *Простота модели.* Следует предпочитать наиболее простую модель при одном и том же уровне точности. Это имеет влияние как на скорость работы модели, так и на её понятность. Кроме того, более простые модели как правило являются более робастными.

**Типы ошибок.** Введём следующие обозначения:

- $a$  — количество «плохих» клиентов, предсказанных верно.
- $b$  — количество «хороших» клиентов, предсказанных верно.
- $c$  — количество «плохих» клиентов, предсказанных как «хорошие».
- $d$  — количество «хороших» клиентов, предсказанных как «плохие».

Точность классификации можно описать одним числом — общей интенсивностью ошибок:

$$ER = \frac{b + c}{a + b + c + d},$$

однако для систем кредитного скоринга необходимо рассматривать отдельно вероятность ошибок первого и второго рода:

$$ER_1 = \frac{c}{a + c}, \quad ER_2 = \frac{b}{b + d}.$$

Число  $ER_1$  характеризует кредитный риск — процент «плохих» клиентов, классифицированных как «хорошие». Значение  $ER_2$  характеризует так называемый *коммерческие риск*, связанный с отказом «хорошим» клиентам.

**Кривая ROC.** В том случае, если скоринговая система на выходе выдаёт непрерывное значение счёта, точность классификации зависит не только от самой модели, но и от порогового значения счёта, начиная с которого принимается положительное решение по выдаче кредита. Для сравнения различных моделей в этом случае применяется *кривая ROC* (receiver operating characteristic), показывающая зависимость  $(1 - ER_1)$  от  $ER_2$ . Чем выше проходит такая кривая, тем точнее классификация независимо от порогового значения (рис. 2).

Применяется также численный показатель, обозначаемый *AUROC* или *AUC* (area under ROC) и равный площади фигуры между кривой ROC и прямой  $1 - ER_1 = ER_2$ . Этот показатель изменяется от 0 (бесполезный классификатор) до  $\frac{1}{2}$  (абсолютно точный классификатор).

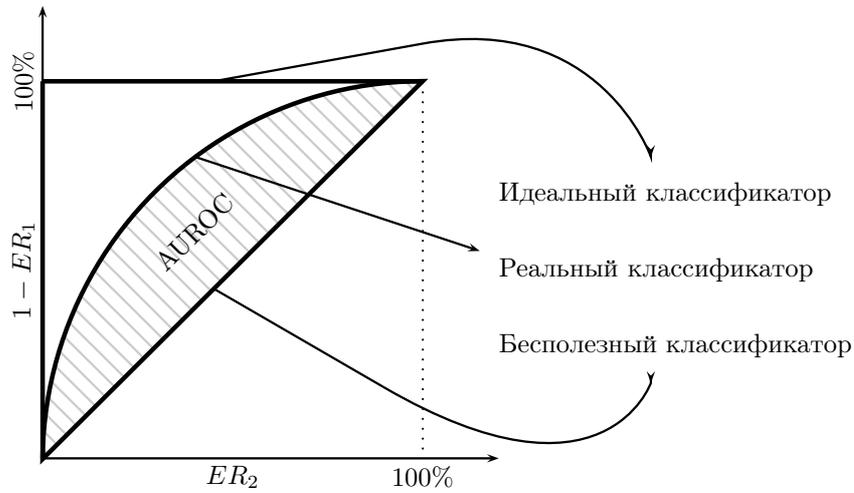


Рис. 2: Кривая ROC

**Матрица штрафов.** Для того, чтобы выбрать пороговое значение счёта, необходимо задаться штрафами за ошибки первого и второго рода (ср. с матрицей штрафов, описанной ранее). Если обозначить последние за  $C_1$  и  $C_2$ , то поиск порогового значения сводится к минимизации ожидаемого штрафа:

$$EC = C_1 \cdot ER_1 \cdot p_B + C_2 \cdot ER_2 \cdot p_G \rightarrow \min.$$

Можно показать, что поиск оптимального порогового значения эквивалентен нахождению касательной к кривой ROC, проходящей под углом  $\arctan(P_G C_2 / P_B C_1)$ .

**Последовательный анализ.** Целью последовательного анализа является сравнение набора классификационных моделей, построенных за счёт изменения размера обучающей выборки. Если модели одинакового качества могут быть получены на различных объёмах данных, следует использовать ту модель, которая получена на меньшей выборке.

Графическое представление результатов последовательного анализа — *кривая обучения*, показывающая зависимость тестовой интенсивности ошибок от размера обучающей выборки или от сложности модели. Результаты экспериментов показывают, что вначале с ростом выборки интенсивность ошибок падает за счёт учёта дополнительной информации. Однако далее интенсивность ошибок стабилизируется и даже может начать возрастать. Это связано с тем, что при больших объёмах данных начинает происходить подгонка под них, либо с тем, что модель достигает предела своих возможностей по извлечению закономерностей из этих данных (рис. 3).

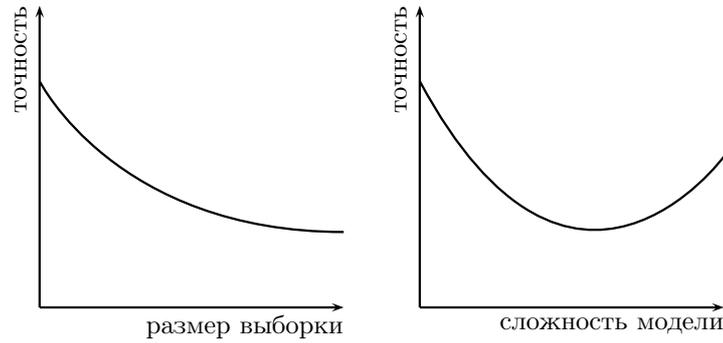


Рис. 3: Кривая обучения

## 2.4 Параметры заявки: дискретные и непрерывные

Среди переменных, описывающих данные кредитной заявки, большую часть обычно составляют дискретные (образование, пол, семейное положение, цель кредита, вид собственности на жилье, род деятельности и т.п.). При этом если некоторые переменные поддаются некоторому упорядочению (например, образование — можно считать, что чем выше уровень, тем больше значение переменной), то для других не существует никакого осмысленного линейного порядка (например, семейное положение или цель кредита). Следовательно, такие переменные нельзя даже приблизительно считать непрерывными, поскольку их значения суть номера ответов на соответствующие вопросы, которые могут располагаться в произвольном порядке.

Если используемая модель скоринга требует использования непрерывных переменных, то можно обойти дискретность переменных, заменив их на большее количество переменных, принимающих значения от 0 до 1. Пусть  $\{X_i\}_{i=1}^n$  — исходный набор (дискретных) переменных, каждая из которых принимает  $r_1, \dots, r_n$  значений соответственно (для упрощения будем считать, что это значения от 1 до  $r_i$ ). Тогда модифицированный набор будет содержать переменные  $\{X_i^j\}$ ,  $i = \overline{1, n}$ ,  $j = \overline{1, r_i}$ , а значения параметров будут определяться по следующему правилу:

$$X_i^j = \delta_{j, X_i} = \begin{cases} 1, & X_i = j, \\ 0, & X_i \neq j. \end{cases}$$

Поскольку значения переменных  $\{X_i^j\}$  уже являются упорядоченными, то по ним можно построить, например, модель линейной регрессии, присвоив каждой из переменных некоторый вес  $w_i^j$ :

$$Y = \sum_{i=1}^n \sum_{j=1}^{r_j} w_i^j X_i^j.$$

Если вернуться к исходному набору переменных  $\{X_i\}$ , то получается, что каждому варианту ответа ( $j$ ) на каждый вопрос ( $i$ ) присвоен некоторый вес ( $w_i^j$ ), а для определения счёта необходимо лишь просуммировать веса выбранных клиентом ответов.

Для некоторых моделей (байесовские сети, деревья классификации), напротив, предпочтительнее использование дискретных, категоризированных данных. В этом случае для каждой непрерывной переменной необходимо разбить область её значений на несколько отрезков, и заменить эту переменную на дискретную, значением которой является номер отрезка, в который попадает значение исходной непрерывной переменной.

Например, если переменная  $X$  есть возраст клиента, то её можно заменить на дискретную переменную  $\bar{X}$ , принимающую следующие значения от 1 до 4, если возраст попадает соответственно на отрезки  $[18, 25]$ ,  $[25, 35]$ ,  $[35, 60]$ ,  $[60, 100]$ .

Выбор параметров дискретизации можно производить из различных соображений: минимизация энтропии (т.е. получение наибольшего возможного количества информации по значению дискретной переменной), равная частота попадания значений переменной в каждый отрезок, равная длина отрезков, или же использование дополнительной (экспертной) информации. Например, указанное выше разбиение для возраста может быть мотивировано использованием информации о жизненном цикле потенциальных заёмщиков.

Часто имеет смысл замена непрерывных переменных на дискретные и затем обратный переход к переменным, принимающим значения 0 и 1. Известно, например, что зависимость кредитоспособности от возраста существенно нелинейна, поэтому применение, скажем, линейной регрессии приведёт к построению неудовлетворительной модели. Однако эту нелинейность можно учесть, если разбить значения возраста на промежутки и присвоить каждому промежутку свой вес. Таким образом, замена непрерывных переменных на дискретные помогает даже в линейных моделях учесть нелинейные зависимости в моделируемых явлениях.

## 2.5 Отсутствующие данные. *Reject Inference*

Многие методы скоринга предполагают, что имеются в наличии *все* значения в историческом наборе данных, используемом при построении модели. Иными словами, в таблице исторических данных не должно быть пропусков. Однако на практике это требование обычно нарушается. Выделяется два основных типа отсутствия данных, которые мы далее рассмотрим.

**Случайно отсутствующие данные.** Исторические данные действительно могут быть с пропусками. Некоторые заявки могли быть заполнены не полностью, или часть информации просто не была введена в базу данных. Другой вариант — данные получены из нескольких источников с немного различающейся структурой заявки. Наконец, в анкету могли быть внесены изменения (добавлен новый вопрос), тогда в данных до этого изменения значение ответа на вопрос будет полностью отсутствовать.

Во всех этих случаях отсутствие данных не представляет существенных проблем при построении модели (кроме, может быть, некоторого ухудшения качества оценки параметров).

Пусть  $X$  — возраст заёмщика,  $Y$  — категория заёмщика,  $Z$  — случайная величина, принимающая значения 1, если есть информация о возрасте, и 0, если данные отсутствуют. В данном случае можно считать, что случайная величина  $Z$  независима с  $X, Y$ , поэтому интересующее нас распределение  $P(Y|X)$  совпадает с распределением  $P(Y|X, Z = 1)$ , которое может быть оценено по имеющимся данным.

Для борьбы со случайно отсутствующими данными могут применяться следующие методы.

- Удаление записей с отсутствующими данными. Если отсутствующих данных относительно немного, то можно просто не учитывать те случаи, в которых есть пропуски данных.
- Замена отсутствующих значений на моду. По имеющимся данным определяется распределение отсутствующего параметра, и затем во всех записях, где он отсутствует, проставляется его мода. Можно использовать как распределение одного параметра, так и условное распределение по присутствующим параметрам. Этот способ также хорошо применять, когда отсутствующих данных относительно мало.
- Моделирование отсутствующих данных. Как и в предыдущем случае, определяется распределение параметра (простое или условное), но только теперь в каждом случае производится случайный эксперимент с использованием этого распределения, и на место отсутствующего значения записывается исход этого эксперимента.
- Замена строки с отсутствующими данными на несколько строк с различными весами. Например, если  $P(X = 1) = 0,2$ ,  $P(X = 2) = 0,8$ , то исходная строка, в которой отсутствовало значение  $X$ , заменяется на две строки: в первой строке  $X = 1$ , во второй —  $X = 2$ , с весами 0,2 и 0,8 соответственно.

**Неслучайно отсутствующие данные.** Ситуация ухудшается, если данные отсутствуют неслучайно. Здесь в свою очередь также возможны два случая.

Во-первых, может отсутствовать часть ответов на вопросы анкеты в случае, если в другой части выбраны определённые параметры. Например, анкета состоит из двух частей, и банк просит заполнять вторую часть клиентов определённого возраста, или клиентов, работающих в определённых отраслях. В этом случае либо воспользоваться дополнительными данными в виде статистики о том, как могли распределены отсутствующие ответы у соответствующих групп клиентов, либо просто построить разные модели для различных групп клиентов.

Однако самая большая проблема — отсутствие данных по клиентам, по тем или иным причинам не получившим кредит. Пусть, как и раньше,  $X$  — один из параметров заёмщика,  $Y$  — наличие дефолта. Кроме того, пусть  $W$  — вся информация о клиенте, не вошедшая в  $X$ . Случайная переменная  $Z$  теперь имеет смысл наличия (1) или отсутствия (0) данных о значении  $Y$ . Введём ещё одну переменную,  $T$ : она равна 1,

если кредит был выдан, и 0, если в кредите было отказано. Можно считать, что  $Z = T$ , то есть имеется информация именно о тех клиентах, которым кредит был выдан, что соответствует реальности — банк не знает, был бы дефолт по клиенту или нет, если кредит не был выдан.

Пусть  $X$  имеет два значения, 0 и 1, и значение 0 однозначно означает высокую вероятность дефолта (например,  $X$  — наличие «определённого места жительства»). Тогда при  $X = 0$  в большинстве случаев будет  $T = 0$ ; только в некоторых случаях  $T = 1$ , когда значения переменных из  $W$  в сумме говорят о достаточной кредитоспособности, несмотря на значение  $X$ .

Если банк имеет информацию только о тех клиентах, которым он выдал кредит, то он может по этим данным оценить вероятность  $p_{T,0} = P(Y = 1|X = 0, T = 1)$ . Однако она не будет иметь никакого отношения к реальной вероятности  $p = P(Y = 1|X = 0)$ . Ничто не мешает обстоятельствам сложиться так, что  $p_{T=0} = 1\%$ , а  $p = 99\%$ . Более того, вероятность  $p_T$  может даже быть меньше  $p_{T,1} = P(Y = 1|X = 1, T = 1)$ , поскольку претенденты с  $X = 0$  должны были пройти более тщательный отбор. Таким образом, может появиться ошибочное предположение, что давать кредит бомжам выгоднее, чем всем остальным.

Построение модели несмотря на отсутствие данных по клиентам, не получившим кредит, в западной литературе носит название *Reject Inference* [12]. Основная проблема заключается в том, что при длительном использовании системы скоринга приходится изменять параметры модели в соответствии со вновь поступающими данными, что приводит к накоплению системной ошибки. Последствия могут быть самые разные: как недооценка риска (см. пример выше), так и переоценка — определённые категории клиентов никогда не получают кредит, у них просто не будет возможности доказать свою кредитоспособность [5].

К настоящему времени показано [12], что распределение  $P(X, W)$  параметров всех потенциальных заёмщиков не даёт дополнительной информации, которая могла бы быть использована при построении модели. Немногие предложенные методы (см. например [6]) основаны на использовании дополнительной информации помимо самих исторических данных.

Если в наличии имеются только исторические данные по кредитам, то приведённый выше пример говорит о том, что при построении новой скоринговой модели по данным о кредитах, выданных по решению старой системы или кредитного офицера, необходимо включить в анкету все те вопросы, которые учитывала старая система (см. [12]).

Одним из способов получить статистику по всем клиентам — принимать на короткое время всех клиентов. Так обычно поступают розничные магазины и фирмы, торгующие по каталогам. Банки, напротив, не могут себе этого позволить из-за высоких потерь при дефолте и вынуждены использовать тот или иной метод *reject inference*.

Для преодоления описанной проблемы можно предложить следующий подход: для клиентов, которым было отказано в кредите, произвести случайные испытания, воспользовавшись данными системы априорного скоринга, и далее использовать исход этих испытаний при построении модели скоринга.

## 2.6 Байесовские классификаторы

Как уже говорилось, байесовские сети позволяют записать многомерное совместное распределение в виде комбинации нескольких распределений меньшей размерности. Более строго, байесовской сетью называется пара  $N = \langle G, \Theta \rangle$ , где  $G$  — ориентированный ациклический граф (ОАГ), а  $\Theta$  — набор условных распределений. Каждая вершина графа соответствует одной из переменных  $X_1, \dots, X_n$ . Для каждой вершины задано условное распределение  $\theta_{X_i|\Pi_{X_i}} = P(X_i|\Pi_{X_i})$ , где  $\Pi_{X_i}$  — множество непосредственных предшественников  $X_i$  в графе  $G$ . Байесовская сеть  $N$  определяет следующее совместное распределение:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i|\Pi_{X_i}).$$

Если известны реализовавшиеся значения одной или нескольких переменных  $X_i = x_i$  ( $i \in I$ , причём возможен случай  $I = \emptyset$ ), то известные алгоритмы вывода по байесовской сети позволяют оценить условные вероятности  $P(X_j|X_I)$  остальных переменных.

В общем случае определение оптимальной структуры сети и вычисление маргинальных вероятностей по имеющимся наблюдениям являются  $NP$ -трудными задачами, поэтому используют приближённые алгоритмы обучения и вывода [21].

Для построения классификатора, основанной на байесовской сети, необходимо вначале найти адекватную структуру графа  $G$ , а затем оценить условные распределения  $\Theta$ . Последняя задача имеет довольно простое решение (при условии отсутствия пропусков в данных), основную же трудность представляет именно поиск оптимальной структуры.

Когда байесовские сети применяются к задаче классификации, граф  $G$  условно разделяется на две части: вершина  $C$ , соответствующая классу клиента, и все остальные вершины. При этом выделяются следующие основные структуры классифицирующей сети [4, 19].

1. *Наивный байесовский классификатор* (naïve Bayes classifier). Из вершины  $C$  проведены стрелки во все входные переменные  $X_1, \dots, X_n$ , и других рёбер у графа  $G$  нет. Обучение классификатора означает оценку условных вероятностей  $P(X_i|C)$ , а классификация производится простым применением формулы Байеса:

$$P(C = c|X = x) = \frac{P(C = c) \prod_{i=1}^n P(X_i = x_i|C = c)}{\sum_{c'} P(C = c') \prod_{i=1}^n P(X_i = x_i|C = c')}.$$

Подобная структура графа означает предположение об условной независимости переменных  $X_i$  при заданном классе  $c$ .

2. *Наивный байесовский классификатор над деревом* (tree augmented naïve Bayes classifier, TAN). Условие условной независимости переменных в  $X$  ослаблено до требования того, чтобы подграф  $G$ , отвечающий переменным из  $X$ , образовывал дерево, то есть у каждой переменной  $X_i$  в качестве непосредственного предшественника помимо класса  $C$  может быть не более одной переменной  $X_j$  (рис. 4).

3. *Наивный байесовский классификатор над сетью* (Bayesian network augmented naïve Bayes classifier). Здесь все переменные  $X_i$  имеют в качестве предшественника класс  $C$ , но при этом они могут образовывать произвольную байесовскую сеть.
4. *Произвольная байесовская сеть*. Не накладывается никаких ограничений на условную независимость  $X$  и  $C$ . При этом вводится понятие *марковского ограждения* (Markov blanket) вершины  $C$ :  $M_C = \Pi_C \cup \Sigma_C \cup \Pi_{\Sigma_C}$ , где  $\Sigma_C$  — множество последователей  $C$  в графе  $G$ . Имеет место условная независимость  $I(C, X \setminus (C \cup M_C) | M_C)$ , поэтому в конечной модели разумно оставлять только вершины из  $C \cup M_C$  (рис. 5).

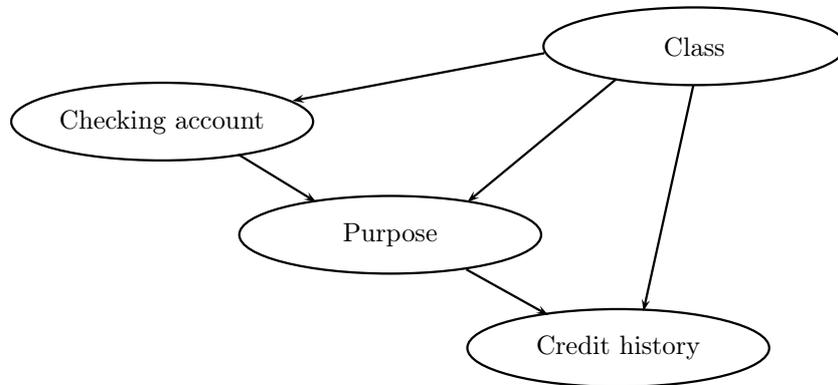


Рис. 4: Структура наивного байесовского классификатора над деревом

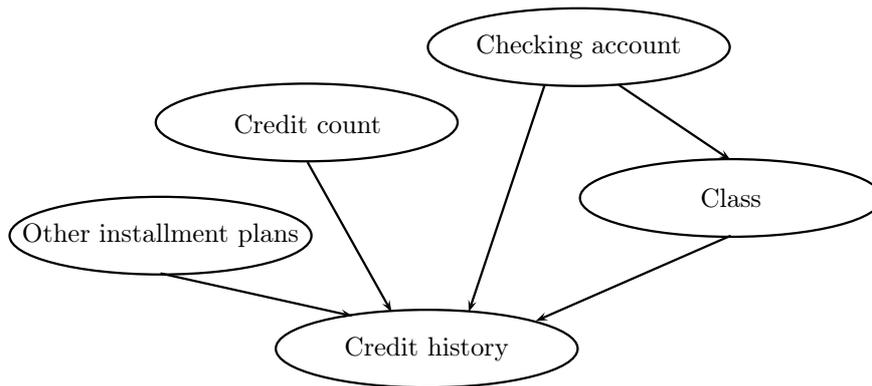


Рис. 5: Структура байесовского классификатора общего вида

Для каждого из перечисленных типов классификаторов существуют свои алгоритмы определения структуры и вывода. В качестве приближённого метода часто

используется МСМС (Markov Chain Monte Carlo) [14, 4], который состоит в последовательной генерации байесовских сетей, причём каждая последующая сеть генерируется марковской цепью с простым распределением вероятностей. После этого можно либо выбрать наиболее вероятную сеть из сгенерированной последовательности, либо применять усреднение по набору наиболее вероятных сетей. Марковское ограждение может быть также построено непосредственно, используя метод, предложенный в [19].

### 3 Поведенческий скоринг

Можно выделить следующие основные модели поведенческого скоринга.

- Методы, аналогичные кредитному скорингу — классификация клиентов в зависимости от их предыдущего поведения (фактически, это тот же кредитный скоринг, только в анкету помимо данных о самом заёмщике добавлено описание его поведения).
- Статистические модели, не ставящие целью объяснить природу того или иного поведения заёмщиков, а лишь аппроксимирующие наблюдаемое поведение. Здесь преобладают модели, основанные на марковских цепях [25, 24, 22, 23]. Последние могут быть использованы как для моделирования поведения одного заёмщика, так и для оценки потерь портфеля кредитов в целом.
- Структурные модели, описывающие динамику кредитоспособности и репутации заёмщика [3, 7]. Доступные статистические данные используются здесь для оценки параметров модели.

#### 3.1 Скоринговые модели

Системы поведенческого скоринга используют информацию о недавнем поведении заёмщика для оценки вероятности дефолта, при этом используется методология, сходная с применяемой в кредитном скоринге. Подготавливается выборка данных по существующим заёмщикам, включающая в себя информацию о транзакциях по их счетам за определённый период времени. Этот период разбивается на две части: период наблюдения и контрольный период, каждый примерно по одному году. На периоде наблюдения вычисляются различные численные характеристики поведения заёмщика: средний, максимальный и минимальный баланс, оборот по счетам, тренды в платежах и балансе и т.д.

Далее каждый заёмщик классифицируется как «плохой» или «хороший» в зависимости от состояния на конец контрольного периода. Далее теми же методами, что и в кредитном скоринге, строится модель, классифицирующая заёмщиков на «хороших» и «плохих».

## 3.2 Модели, основанные на марковских цепях

Марковские модели основаны на двух предположениях: во-первых, набор состояний действительно описывает всевозможные состояния, в которых может находиться заёмщик, и во-вторых, динамика изменения состояний действительно удовлетворяет марковскому свойству.

В качестве состояний обычно выбираются различные уровни задержки платежей, например, количество месяцев задержки. Переходные вероятности определяются по историческим данным.

Для построения более адекватной модели в состояние могут включаться дополнительные данные (о состоянии счёта, размере долга и т.п.). Другой подход состоит в сегментировании заёмщиков и последующем определении состояний для каждого сегмента.

Далее необходимо проверить справедливость марковского свойства для построенной модели. Для этого вычисляются частоты перехода  $i \rightarrow j$  с частотами переходов  $i \rightarrow k \rightarrow j$  для любого состояния  $k$ , и затем по критерию  $\chi^2$  проверяется гипотеза о совпадении этих распределений.

Даже если при выборе состояний использовалась дополнительная информация о клиентах или была произведена сегментация, марковское свойство может всё равно не выполняться. В таких случаях следует либо вернуться и определить другие состояния, либо рассмотреть цепь Маркова второго и более высокого порядков. Иными словами, в качестве нового состояния выбирается пара из текущего состояния и состояния на предыдущем периоде. Это существенно увеличивает число состояний и возможных переходов, но за счёт того, что многие переходы невозможны, матрица переходных вероятностей будет разреженной.

Если и после этого марковское свойство нельзя считать выполненным, это может означать нестационарность моделируемого процесса. В этом случае состояние может зависеть от

- текущего времени  $t$ ;
- количества времени, в течение которого заёмщик является клиентом банка ( $s$ );
- различные экономические показатели, например, базовая процентная ставка  $r$ .

Далее необходимо оценить переходные вероятности  $p_{i,j}(r, s, t)$ , т.е. вероятность перемещения заёмщика из состояния  $i$  в состояние  $j$ , когда описанные выше параметры имеют значения  $r, s, t$ . Одной из реализованных на таком пути моделей была оценка переходных вероятностей  $p_{ij}(r, t)$ :

$$p_{ij}(r, t) = p_{ij}^0 + a_{ij}t + b_{ij}r,$$
$$\sum_j p_{ij}^0 = 1, \quad p_{ij}^0 \geq 0, \quad \sum_j a_{ij} = 0, \quad \sum_j b_{ij} = 0, \quad p_{ij}(r, t) \geq 0.$$

При этом значениям коэффициентов  $a_{ij}, b_{ij}$  можно придать интерпретацию в терминах факторов, вызывающих задержки платежей.

### 3.2.1 Определение состояний: анализ пуассоновских смесей

Одним из подходов к выделению основных состояний заёмщиков является идентификация параметров смесей распределений, и в частности — пуассоновских смесей [11, 8, 20].

Предположим, что количество пропущенных платежей для каждого клиента является пуассоновским процессом с интенсивностью  $\lambda_i$ . Значение  $\lambda_i$  можно было бы использовать в качестве одной из составляющих, описывающих состояние клиента, если бы только можно было оценить её с достаточной достоверностью. Оценкой максимального правдоподобия является

$$\hat{\lambda}_i = \frac{\text{число пропущенных платежей}}{\text{общее число платежей}}.$$

Однако такая оценка имеет слишком низкую статистическую значимость. Например, если клиент не пропустил ни одного платежа за 12 месяцев, это ещё не означает, что у него  $\lambda_i = 0$ . Если взять больший промежуток времени, то оценка становится более точной, но при этом нельзя забывать, что во-первых  $\lambda_i$  меняется со временем (это состояние заёмщика), а во-вторых, если конечной целью является построение марковской цепи и вычисление переходных вероятностей, то использовать большие промежутки времени для оценки  $\lambda_i$  не представляется возможным.

Пусть имеется портфель из  $N$  заёмщиков, тогда общим распределение числа неплатежей за единичный период времени является

$$P(x = k) = \frac{1}{N} \sum_{i=1}^N \frac{e^{-\lambda_i} \lambda_i^k}{k!}.$$

Сделаем упрощающее предположение: заёмщиков можно разделить на  $n$  групп численностью  $n_1, \dots, N_n$ , так что каждая группа имеет одинаковую интенсивность неплатежей  $\lambda_i$ . Тогда

$$P(x = k) = \sum_{i=1}^n \frac{e^{-\lambda_i} \lambda_i^k}{k!} p_i,$$

где  $p_i = \frac{N_i}{N}$  — доля группы  $i$  в общей численности клиентов.

Зафиксируем теперь достаточно большой период времени и вычислим эмпирическое распределение числа неплатежей за этот период. После этого вычислим оценку максимального правдоподобия, т.е. найдём число  $\hat{n}$ , частоты интенсивности неплатежей  $\hat{\lambda}_i$  и смешивающее распределение  $\hat{p}_i$ . Тогда в качестве состояния заёмщика можно использовать номер  $i$ , отражающий его текущую интенсивность неплатежей.

Пусть по клиенту имеется информация о количестве неплатежей  $k$  за период времени  $T$ . Тогда по формуле Байеса получаем следующее апостериорное распределение состояния:

$$P(C = i | x(T) = k) = \frac{P(x(T) = k | C = i) P(C = i)}{P(x(T) = k)} = \frac{e^{-\hat{\lambda}_i T} (\hat{\lambda}_i T)^k \hat{p}_i}{\sum_{j=1}^{\hat{n}} e^{-\hat{\lambda}_j T} (\hat{\lambda}_j T)^k \hat{p}_j}$$

Когда доступны данные за очень короткий период, вероятности  $P(C = i|x(T) = k)$  близки к априорным вероятностям  $\hat{p}_i$

$$\lim_{\alpha \downarrow 0} P(C = i|x(\alpha T) = \alpha k) = \hat{p}_i.$$

Наоборот, при наличии данных за большой период времени ожидаемая интенсивность дефолтов приближается к одному из значений  $\hat{\lambda}_i$ .

Для определения состояния конкретного заёмщика можно выбрать либо состояние с максимальной апостериорной вероятностью, либо взять в качестве ожидаемой интенсивности дефолта

$$\lambda = \sum_{i=1}^n \hat{\lambda}_i P(C = i|x(T) = k).$$

Отметим, что если все  $\hat{\lambda}_i > 0$ , то  $\lambda > 0$ , даже если  $k = 0$ , то есть клиент исправно исполняет все платежи. Это позволяет учесть возможность неплатежа у клиента, который до сих пор не пропустил ни одного платежа. С другой стороны,  $\lambda \rightarrow 0$  при  $T \rightarrow \infty$ , то есть с течением времени клиент, не пропускающий платежи, считается всё более и более надёжным.

Опишем теперь, как можно построить марковскую цепь, используя полученные результаты. Пусть  $\hat{n}$ ,  $\hat{\lambda}_i$ ,  $\hat{p}_i$  — параметры модели, оцененные по данным на достаточно длинном промежутке времени. Зафиксируем теперь два последовательных промежутка времени, скажем, длиной в год (примем это за единичный отрезок времени). По каждому клиенту  $m = \overline{1, N}$  вычислим количество платежей, пропущенных в первом и втором периоде ( $k'_m$  и  $k''_m$  соответственно). Определим состояния цепи как числа  $1, \dots, \hat{n}$  и вычислим оценки переходных вероятностей

$$\hat{p}(i \rightarrow j) = \frac{\sum_{m=1}^N P(C = i|x(1) = k'_m) P(C = j|x(1) = k''_m)}{\sum_{j'=1}^{\hat{n}} \sum_{m=1}^N P(C = i|x(1) = k'_m) P(C = j'|x(1) = k''_m)}$$

Очевидно, что тот же самый подход можно применить, если состояния в цепи имеют более сложный вид, включая в себя не только интенсивность неплатежей, но и другую информацию: состояние счёта, сумму непогашенного долга и т.п.

### 3.3 Структурные модели

В работе [3] описана структурная модель кредитного риска частных лиц с использованием теории оценки опционов и понятия репутации заёмщика. Для этого делаются следующие предположения:

- Существует ненаблюдаемая случайная величина  $Q_i$ , называемая кредитоспособностью заёмщика  $i$ , заключающая в себе всю информацию, необходимую для оценки кредитного риска.

- Хотя нет возможности наблюдать  $Q_i$  непосредственно, рынок использует доступную информацию о заёмщике, такую как успешный возврат предыдущих кредитов, финансовое состояние заёмщика, другие факторы, для того, чтобы оценить  $Q_i$  или построить некий суррогат этой величины.
- Вероятность того, что данному клиенту будет выдан кредит, является строго возрастающей функцией  $Q_i$ :  $P_{cr} = f(Q_i)$ .
- Существуют механизмы передачи информации о дефолте произвольной кредитной организации. Если у заёмщика происходит дефолт, об этом узнают все кредитные организации и у него не будет доступа к кредиту в ближайшее время. Это называется «потерей репутации».
- Доступ к кредиту представляет ценность для заёмщика. Эта ценность, выраженная в денежном эквиваленте, называется репутацией заёмщика,  $R_i$ . Это строго возрастающая функция  $P_{cr}$ ,  $R = h(P_{cr})$ .
- Если заёмщик не может внести очередной платёж в срок, он возьмёт дополнительный кредит на соответствующую сумму.

При указанных предположениях можно использовать для оценки кредитного риска те же рассуждения, что и в теории оценки опционов Мертона. У заёмщика есть опцион колл на его репутацию с ценой исполнения, равной сумме долга. Как только ценность репутации  $R_i$  опускается ниже суммы долга, заёмщик объявляет дефолт. Из монотонности функций  $f(\cdot)$ ,  $g(\cdot)$  вытекает, что дефолт происходит при падении кредитоспособности  $Q_i$  ниже определённого уровня,  $K_{Q_i} = K_{Q_i}(i, D_i)$ .

Изменение  $Q_i$  с течением времени описывается стохастическим дифференциальным уравнением

$$dQ_i(t) = \mu dt + \sigma dW(t) + a(t) dY(t),$$

где  $\mu$  — параметр дрейфа,  $\sigma$  — волатильность,  $W(t)$  — стандартное броуновское движение.  $a(t)$  — размер скачка в момент  $t$  с распределением  $N(\mu_a, \sigma_a^2)$ ,  $Y(t)$  — пуассоновский процесс с интенсивностью  $\lambda$ . Случайные величины  $a(t)$ ,  $dW(t)$ ,  $dY(t)$  предполагаются взаимно независимыми.

Поскольку величина  $Q_i$  не является наблюдаемой, то она заменяется в этом уравнении на суррогатное значение — текущие кредитный счёт  $S_i$ . Параметры модели определяются по историческим данным с помощью метода Markov Chain Monte Carlo (MCMC) [14].

## 4 Заключение

В качестве примера сочетания ряда приведённых методов в цельной системе скоринга можно предложить следующую составную модель:

- По историческим данным о неплатежах клиентов определяются параметры  $\hat{n}$ ,  $\hat{\lambda}_i$ ,  $\hat{p}_i$  пуассоновской смеси, описывающей интенсивность неплатежей.
- Определяются переходные вероятности  $\hat{p}(i \rightarrow j)$  переключения с интенсивности неплатежей  $\hat{\lambda}_i$  на  $\hat{\lambda}_j$ .
- Для каждого из существующих клиентов определяется распределение вероятностей по состояниям  $1, \dots, \hat{n}$ , соответствующим различным интенсивностям неплатежей. Эти данные используются для построения байесовской сети классификации для кредитного скоринга новых клиентов.
- Для каждого нового клиента с помощью построенной байесовской сети вычисляется распределение по состояниям  $\hat{\lambda}_i$ . Для уточнения распределения может использоваться дополнительная информация, например, система априорного скоринга. Вычисленное распределение позволяет спрогнозировать вероятность дефолта и график погашения долга, поэтому может быть использовано для принятия решения о выдаче кредита или отказе.
- По мере поступления информации о количестве неплатежей клиента по формуле Байеса уточняется его распределение по состояниям  $\hat{\lambda}_i$ . При этом в качестве априорных вероятностей используются вероятности, выданные системой кредитного скоринга.
- Используя матрицу переходных вероятностей, на уровне портфеля прогнозируются ожидаемые потери от дефолтов.

## Список литературы

- [1] *Соложенцев Е. Д.* Сценарное логико-вероятностное управление риском в бизнесе и технике. СПб: Бизнес-пресса, 2004.
- [2] Machine Learning Repository (University of California, Irvine). <http://www.ics.uci.edu/~mlern/MLSummary.html>.
- [3] *de Andrade F. W. M., Thomas L.* Structural models in consumer credit. Working paper, July 2004.
- [4] *Baesens B., Egmont-Petersen M., Castelo R., Vanthienen J.* Learning bayesian network classifiers for credit scoring using Markov chain Monte Carlo search // Proc. International Congress on Pattern Recognition. 2002.
- [5] *Bridges S., Disney R.* Modelling consumer credit and default: The research agenda. Working paper, Experian Centre for Economic Modelling, December 2001.

- [6] *Chen G. G., Åstebro T.* Bound and collapse bayesian reject inference when data are missing not at random // “Mathematical Approaches to Credit Risk Management”, Conference Proceedings, Banff International Research Station for Mathematical Innovation and Discovery / Eds. Åstebro T., Beling P., Hand D., Oliver B., Thomas L. B. October 2003.
- [7] *De Giorgi E., Burkhard J.* An intensity based non-parametric default model for residential mortgage portfolios. Working paper, University of Zurich, May 2004.
- [8] *Dellaportas P., Karlis D., Xekalaki E.* Credit scoring via bayesian analysis of finite poisson mixtures. Working paper, Athens University of Economics and Business, Department of Statistics, 1999.
- [9] *Durand D.* Risk elements in consumer installment financing. NY: National Bureau of Economic Research, 1941.
- [10] *Fisher R. A.* The use of multiple measurements in taxonomic problems // *Annals of Eugenics*. 1936. N. 7. p. 179–188.
- [11] *Guillén M., Artís M.* Count data models for a credit scoring system. Paper presented at the Third Meeting on the European Conference Series in Quantitative Economics and Econometrics on Econometrics of Duration, Count and Transition Models, December 1992.
- [12] *Hand D. J., Henley W. E.* Can reject inference ever work? // *IMA Journal of Mathematics Applied in Business and Industry*. 1993. N. 5. p. 45–55.
- [13] *Hand D. J., Henley W. E.* Statistical classification methods in consumer credit // *Journal of the Royal Statistical Society, Series A*. 1997. V. 160. p. 523–541.
- [14] *Johannes M., Polson N.* MCMC methods for continuous-time financial econometrics. Prepared for the Handbook of Financial Econometrics, 2003. <http://www-1.gsb.columbia.edu/faculty/mjohannes/>.
- [15] *Liu Y.* New issues in credit scoring application. Arbeitsbericht 16/2001, Institut für Wirtschaftsinformatik, 2001.
- [16] *Liu Y.* A framework of data mining application process for credit scoring. Arbeitsbericht 01/2002, Institut für Wirtschaftsinformatik, 2002.
- [17] *Liu Y.* The evaluation of classification models for credit scoring. Arbeitsbericht 02/2002, Institut für Wirtschaftsinformatik, 2002.
- [18] *Lovie A. D., Lovie P.* The flat maximum effect and linear scoring models for prediction // *Journal of Forecasting*. 1986. N. 5. p. 159–186.

- [19] *Madden M. G.* Evaluation of the performance of the Markov blanket Bayesian classifier algorithm. Technical Report NUIG-IT-011002, Department of Information Technology, National University of Ireland, Galway, 2002.
- [20] *Mnatsakanov R. M., Klaasen C. A. J.* Estimation of the mixing distribution in Poisson mixture models: uncensored and censored samples // Proceedings of the Hawaii International Conference on Statistics and Related Fields. Honolulu, Hawaii, USA, June 2003.
- [21] *Neapolitan R. E.* Learning Bayesian Networks. NJ: Prentice Hall, 2003.
- [22] *Scallan G.* Markov models: an introduction. a new approach to bad debt modeling. Technical report, ScorePlus Research, June 1990.
- [23] *Scallan G.* Bad debt projection models: an overview of modeling approaches. Technical report, ScorePlus Research, February 1998.
- [24] *Thomas L. C., Ho J., Scherer W. T.* Time will tell: behavioural scoring and the dynamics of consumer credit assessment // IMA Journal of Management Mathematics. 2001. V. 12. p. 89–103.
- [25] *Thomas L. C.* A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers // International Journal of Forecasting. 2000. V. 16. p. 149–172.