Рындина С.В. Математические модели application-скоринга. // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей VIII Всерос. научно-техн. конф.– Пенза: ПДЗ, 2008. – С. 32-35.

МАТЕМАТИЧЕСКИЕ МОДЕЛИ APPLICATION-СКОРИНГА

С.В. Рындина

Пензенский государственный педагогический университет
им. В.Г. Белинского,
г. Пенза

Система кредитного скоринга – это система управления риском, связанным с предоставлением кредита. Прежде всего это касается кредитования физических лиц и связано с массовостью предоставления подобных кредитов и возможностью выделения групп заемщиков, достаточно однородных по своим социально-финансовым показателям.

Скоринговая система, как правило, включает в себя несколько типов скоринга, которые позволяют управлять всем бизнес-процессом кредитования. На этапе выдачи кредита это application-скоринг (оценка претендента на предоставление кредита с точки зрения его возврата) и fraud скоринг (оценка вероятности, что потенциальный заемщик – мошенник). При обслуживании кредитов – это collection-скоринг (управление проблемной задолженностью), behavioral-скоринг (оценка поведения заемщика при погашении кредита). Скоринговая система решает две основные задачи: автоматизирует процесс принятия решения о предоставлении кредита и последующее управление его погашением, а также создает на основе накопленной информации о заемщиках и их поведении при погашении кредита скоринговые карты.

Основа application скоринга – это скоринговая, оценочная карта, которая позволяет на основе анкетных данных рассчитать балл аппликанта (претендента) или вероятность погашения им суммы кредита, на которую он претендует. На основании заложенных уровней риска (максимальная вероятность невозврата кредита для заемщика, процент потерь банка для этого вида деятельности и т.п.) принимается решение о возможности предоставления кредита очередному заемщику. Скоринговая карта может разрабатываться сторонней организацией для конкретного банковского продукта. Тогда она скорее всего базируется на экспертных оценках кредитных аналитиков и представляет собой балльную систему оценки заемщика. На основе допустимого уровня потерь или процента невозвратов определяется балл отсечения, ниже которого принимается решение очередному заемщику отказать.

Более перспективными представляются динамические скоринговые карты. В их основе математическая модель, которая на основе имеющейся информации о «хороших», т.е. выплаченных кредитах, и «плохих», т.е. просроченных и невозвращенных, оценивает вероятность возврата кредита потенциальным заемщиком. Такая модель не только позволяет локализовать риски, разбивая всех претендентов на группы, для которых можно определить вероятность погашения кредита, дисциплину погашения, максимальную сумму кредита, вероятность мошенничества, но и динамично подстраивается под изменяющиеся условия. Так, с течением времени изменения на рынке труда могли привести к повышению или снижению уровня заработной платы для определенной социальной группы. Анализируя новую информацию, система перенастраивается для определения новых границ групп заемщиков с устойчивыми финансовыми показателями и пересчитывает значения этих показателей.

Application-скоринг может быть построен на основе различных математических моделей (или их сочетаний). Критерием выбора модели для использования может быть тестирование правильности предсказания по выборочным данным, для которых результат взаимодействия с банком уже известен.

Во многих скоринговых системах для классификации клиентов используется множественная регрессия, а в последнее время логистическая регрессия.

Регрессионный анализ позволяет получить значимый прогноз при условии наличия в выборке примерно равного количества «хороших» и «плохих» кредитов. Причем построенная модель не будет играть роль «черного ящика»: она позволит проанализировать, какие характеристики клиентов (экзогенные переменные, значения которых задаются извне) наиболее весомы для классификации клиента как «хорошего». Однако множественная регрессия в задаче классификации имеет тот существенный недостаток, что на результат (эндогенную переменную, формируемую внутри модели) не накладывается ограничений. Так как основная цель построенной модели – классифицировать потенциального заемщика, то при наличии, например, двух групп: «хорошие» и «плохие», значений, которые может принимать эндогенная переменная, всего два. Значения, которые принимает зависимая переменная, можно интерпретировать как вероятность возвращения кредита: 0 – кредит невозвращен, 1 – кредит возвращен. В этом случае должна быть построена модель бинарного выбора. Однако выбор может быть и множественным, если рассматривать зависимую переменную *y* как несколько альтернатив, которые некоторым образом упорядочены. Например: 1) вернул досрочно, 2) вернул в срок, без штрафов, 3) вернул в срок, но были просрочки платежей, 3) вернул с опозданием, заплатив все штрафы и комиссии, 4) не вернул, личный дефолт, … В этом случае рейтинг показывает, насколько выгоден тот или иной клиент для банка.

Для прогноза зависимой переменной в случае бинарного выбора можно построить линейную модель вероятности, *probit*- и *logit*-модели.

Линейная модель вероятности:

$$y_i = x_i^T \beta + \varepsilon_i\,,\ i = \overline{1, n}\,, \tag{1}$$

где i – номер наблюдения; $y_i = P(y_i = 1)$ – вероятность возврата кредита; $\beta = (\beta_1, \ldots, \beta_k)^T$ – набор неизвестных параметров (описывают силу влияния соответствующего фактора на результат); $x = (x_1, \ldots, x_k)^T$ – набор объясняющих факторов (социально-финансовые характеристики потенциальных заемщиков); ε – случайная ошибка, влияние неучтенных факторов.

Недостаток линейной модели вероятности в том, что прогнозные значения, рассчитанные по оцененной модели, могут оказаться за пределами отрезка $[0, 1]$. В этом случае интерпретировать их невозможно. При большом числе наблюдений и при первичной обработке данных недостатки модели несущественны.

Для *probit*- и *logit*-модели в правой части уравнения (1) рассматривают некоторую функцию $F(x_i^T\beta)$, область значений которой лежит в отрезке $[0,1]$.

Пусть $u = \frac{x_i^T\beta}{\sigma}$ (σ – среднее квадратическое отклонение для ε). Если в качестве функции F используется функция стандартного нормального распределения

$$F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{z^2}{2}} dz,$$

то получаем *probit*-модель.

Если функция логистического распределения

$$F(u) = \frac{e^u}{1+e^u},$$

то получаем *logit*-модель.

Оценки параметров *probit*- и *logit*-моделей обычно получают с помощью метода максимального правдоподобия. Метод наименьших квадратов – традиционный инструмент оценки параметров регрессионной модели – неприменим, так как в этих моделях нарушены предположения о свойствах случайного члена ε. Выбор обычно делается в пользу той модели, для которой функция правдоподобия дает большее значение.

Полученный результат для *probit*- и *logit*-моделей интерпретируется как вероятность возвращения кредита потенциальным заемщиком (и значения не выйдут за пределы отрезка $[0,1]$). Однако для этих моделей сталкиваемся с затрудненной интерпретацией параметров $\beta = (\beta_1, \ldots, \beta_k)^T$, эффект влияния каждого фактора становится величиной переменной, зависящей от значений других факторов.