УДК 330.43, 519.2, 519.86

Сорокин Александр Сергеевич
ФГБОУ ВПО «Московский государственный университет экономики, статистики и информатики»
Россия, Москва[1]
Доцент кафедры Математической статистики и эконометрики
Московский финансово-промышленный университет «Синергия»
Россия, Москва[2]
Доцент кафедры Бизнес-статистики
Кандидат экономических наук
E-Mail: alsorokin@mail.ru

Построение скоринговых карт с использованием модели логистической регрессии

Аннотация: В банковской сфере при управлении кредитными рисками одна из ключевых задач — оценка кредитоспособности заемщиков. Результаты оценки индивидуальных рисков являются основой для анализа рисков всего кредитного портфеля. Оценка риска невозврата кредита по конкретному заемщику на практике осуществляется в рамках двух основных подходов — на основе субъективного заключения экспертов или на основе автоматизированных систем скоринга.

В основе построения скоринговой системы могут браться различные статистические модели. Эти модели могут быть получены методами линейной регрессии, логистической регрессии, дискриминантного анализа, деревьев решений, нейронных сетей и др. Однако логистическая регрессия является наиболее часто используемой на практике математической моделью для построения скоринговой карты. Настоящая работа посвящена рассмотрению различных подходов и методик к построению скоринговых карт на базе логистической регрессии, а также проблемам, которые могут возникать при построении скоринговых моделей.

В статье рассматривается методика эконометрического моделирования вероятности дефолта по кредитам на основе модели логистической регрессии. Акцентируется внимание на методологических аспектах построения модели. Основные проблемы построения модели иллюстрируются практическими расчетами. Показывается методика перевода полученных коэффициентов модели логистической регрессии в скоринговую карту. Приводится пример построения скоринговой карты.

Авторские выводы и рекомендации могут быть использованы специалистами по управлению рисками в коммерческих банках при построении скоринговых систем и проверки их работы.

Ключевые слова: Кредитный риск; кредитный скоринг; логистическая регрессия; коммерческий банк; управление рисками; скоринговые карты; категоризация количественных переменных; информационное значение, вес категорий предикторов; валидация модели.

Идентификационный номер статьи в журнале 180EVN214

[1] 119501, г. Москва, ул. Нежинская, 7, МЭСИ, кафедра Математической статистики и эконометрики

[2] 125190, г. Москва, Ленинградский пр-кт, д. 80, МФПУ «Синергия», кафедра Бизнес-статистики

1. Введение

Наибольшее распространение в банковской сфере получил кредитный скоринг. Кредитный скоринг[3] можно опередить как метод начисления потенциальным заемщикам определенного количества баллов на основе информации о его социально-демографическом положении, кредитной истории, параметрах запрашиваемого кредита, и принятие решения о выдаче или об отказе в кредите на основе набранного суммарного количества баллов. На настоящий момент банки предъявляют повышенные требования к риск-аналитике в связи с участившимися случаями мошенничества и ростом числа невозвратных кредитов. По данным Национального бюро кредитных историй по состоянию на 1 января 2014 года потери кредиторов от мошенников составили 153 млрд руб., тогда как годом ранее их объем был 67 млрд руб[4]. На практике возникает задача не только принятия решения в отказе или выдачи кредита конкретному заемщику на основе набранного количества баллов, но и задача определения оптимального минимального количества набранных баллов для выдачи кредита. Вторая задача решается на основе анализа распределения баллов «надежных» и «ненадежных» заемщиков на основе полученной скоринговой карты и тесна связана с анализом соотношения риска и доходности во всем кредитном портфеле банка. Таким образом, кредитный скоринг является инструментом снижения рисков невозврата кредитов, а также помогает определить оптимальную структуру кредитного портфеля, корректировать процентные ставки по кредитам в зависимости от уровня риска.

В большинстве коммерческих банков скоринговые модели являются собственными разработками с различными методиками на основе данных о заемщиках конкретного банка прошлых лет, или являются готовыми решениями специализированных фирм на основе данных о заемщиках нескольких банков или финансовых институтов[5]. И в первом и втором случае методики построения скоринговых карт, как правило, составляют коммерческую тайну. Методы построения скоринговых моделей и на их основе скоринговых карт разбираются в таких работах как (Naeem, 2006); (Lewis, 1992); (Allison, 1999); (Scallan 1999); (Anderson, 2007). Обзор практических статей по кредитном скорингу содержит работа (Mays et al., 2001).

2. Подготовка данных для построения скоринговой модели

2.1. Исходная информационная база

В основе построения скоринговых карт лежат статистические модели. Для их построения должна быть достаточная и качественная информация о заемщиках банка. Качество исходных статистических данных для построения статистической модели определяет ее точность прогнозирования и успех разработки скоринговой системы в целом.

Разработка скоринговой модели строится на анализе предыдущего кредитного опыта. Достаточный объем информации — это одна из главных предпосылок построения модели. Количество данных может варьироваться в зависимости от конкретных моделей, но в целом данные должны удовлетворять требованиям статистической значимости и случайности. Исходные данные для построения модели могут содержать внутренние данные анкет

[3] Этот вид скоринга называют еще скорингом заявок (от англ. application-scoring).

[4] По данным ОАО «НБКИ», см. http://www.nbki.ru/press/pressrelease/?id=2152.

[5] Например, такие услуги предоставляют компании Equifax, Experian, Scorto, FICO и др.

заемщиков банка, а также внешние данные кредитных историй, содержащие сотни тысяч записей[6].

В идеале модели скоринга должны применятся в отношении тех же кредитных продуктов, сектора рынка, и экономической ситуации, которые легли в основу данных о прошлом кредитном опыте. Например, сведения по потребительским кредитам не могут адекватно использоваться при разработке скоринговой карты по автокредитованию. Для построения точной скоринговой модели исходные данные должны обладать определенность исторической давностью. Это требование определяет период, за который собираются данные. Например, данные по потребительским кредитам, одобренным 3 месяца назад, не подойдут для построения модели, одобренным 3 года назад скорее подойдут, а 10 лет назад будут являться достаточно устаревшими. Исторический период данных для построения модели определяется, как правило, видом скоринга и видом кредитования, а также требованиям надзорных органов.[7]

Данные об определенном типе клиентов необходимо исключить из исходной информационной базы. Это могут быть нетипичные клиенты — мошенники, сотрудники банка, VIP клиенты, умершие клиенты, несовершеннолетние, двойные заявки, кредиты по украденным картам и др. Также из базы должны быть исключены кредиты с аномально большими суммами кредитов, нестандартными условиям погашения, нетипичными целями займа. Дополнительным критерием отбора данных может служить вид кредитования или регион рынка, для которого строиться скоринговая карта.

Иногда использование нескольких скоринговых карт для одного портфеля по виду кредитования обеспечивает лучшее дифференцирование риска, чем использование одной скоринговой карты. Для реализации этого подхода часто перед построением скоринговой модели исходную базу клиентов сегментируют с помощью многомерных статистических методов, например, кластерного анализа, деревьев решений или эвристическими методами.

2.2. Определение зависимой переменной

Выбор зависимой переменной определяется целью построения скоринговой модели. Цели могут быть общими, например, сокращение потерь по новым кредитным счетам, и конкретными, например, сокращение числа дефолтов по одобренным заявкам в течение 3-х месяцев после принятия положительного решения. Зависимая переменная может принимать количественные и качественные значения. Примером количественной целевой переменной является средняя сумма, которую погасит заемщик по просроченному кредиту. В скоринге заявок зависимая переменная принимает категориальную шкалу измерения.

На этапе определения зависимой переменной заемщиков делят на три группы: «плохие», «хорошие» и «неопределенные». Для мошенников, банкротов и безнадежных кредитов критерий определения «плохого» заемщика однозначен. В отношении остальных заемщиков банка критерием определения «плохого клиента» является, как правило, количество дней просрочки платежа по кредитам. К группе «неопределенных» клиентов могут относить клиентов с недостаточной кредитной историей, имеющие небольшую допустимую просрочку платежа и др. При построении скоринговой карты используются только клиенты, определенные

[6] Примером такой базы могут служить, например, данные трех кредитных Бюро, содержащие информацию о 50 тыс. заемщиках в 28 переменных почти по 300 тыс. кредитам на сайте https://www.tcsbank.ru/tournament/.

[7] Например, для скоринга заявок потребительских кредитов обычно это данные за последние 2–5 лет, для поведенческого скоринга — 6–12 месяцев.

как «плохие» и «хорошие». Доля «неопределенных» клиентов учитывается при построении окончательного прогноза зависимой переменной по модели логистической регрессии. Также при построении скоринговых моделей часто в отдельную категорию выделяют «отклоненных» клиентов, т.е. заемщиков, которым отказали в выдаче кредита. Учет «неопределенных» и «отклоненных» заемщиков позволяет учесть в обучающей выборке данных для построения модели пропорции генеральной совокупности заемщиков[8]. Процесс построения скоринговой модели часто разбивают на два этапа: 1) построение первичной модели по данным «плохих» и «хороших» клиентов без учета «отклоненных» клиентов; 2) построение конечной модели с учетом анализа отклоненных заявок. Многие эксперты в области кредитного скоринга отмечают, что анализ отклоненных заявок клиентов требует больших ресурсов и не всегда приводит к улучшению качества конечной скоринговой модели[9].

Наиболее часто используемый вид измерения зависимой переменной — категориальный с двумя категориями. Обычно к категории «плохой» относят клиентов, имеющих просроченную задолженность 90 дней и более[10]. Для моделирования значений такой переменной идеально подходит логистическая регрессия. Банк может строить различные скоринговые карты с разными значениями зависимой переменной, вводя дополнительные критерии определения «плохого» и «хорошего» заемщика, а также меняя срок просрочки платежей. Примерами зависимой переменной могут быть наличие просроченной задолженности более 30 дней, 60 дней, 90 дней и более по одному кредиту на текущий момент или худший статус за все время кредитной истории, размер просроченной задолженности, количество просрочек более заданного числа дней и др.

2.3. Определение независимых переменных

В качестве независимых переменных при построении скоринговой модели могут быть данные из кредитной заявки: социально-демографические данные о заемщике (пол, семейное положение, возраст, должность, общий стаж работы и стаж работы на последнем месте, срок проживания по текущему адресу, наличие детей, уровень образования, доход заемщика и доход семьи, работает или на пенсии и др.); информация о запрашиваемом кредите (срок погашения кредита по договору, сумма кредита, размер ежемесячных платежей, размер первоначального взноса, цель кредита, наличие обеспечения и др.); реже используются маркетинговые данные (источник рекламы, проводимая маркетинговая программа, мотив выбора банка).

Следующий тип данных для формирования независимых переменных — внутренняя кредитная история заемщика: текущей баланс счета, задолженность на данный момент, количество счетов, наличие и объем сбережений, число предыдущих кредитов в банке, наибольшее значение суммы задолженности по прежним кредитным счетам, наличие просроченных платежей, наличие других банковских продуктов и услуг, регулярность выплат

[8] Данный аспект и анализ «отклоненных» заявок не рассматриваются в настоящей работе. Подробнее см.: (Naeem, 2006).

[9] Например, см.: (Ковалев, Корженевская, 2008).

[10] Этот период определяется требованиями банковского надзора. В соответствие с соглашением Базель II дефолт должника считается произошедшим, когда имело место одно или оба из следующих событий: банк считает, что должник не в состоянии полностью погасить свои кредитные обязательства перед банком без принятия банком таких мер, как реализация обеспечения (если таковое имеется); должник более чем на 90 дней просрочил погашение любых существенных кредитных обязательств перед банком.

прежних долгов по всем обязательствам и др. Для новых клиентов банка эта информация недоступна.

Одним из основных источников данных для формирования независимых переменных в скоринговой модели являются данные Бюро кредитных историй на момент подачи заявки заемщиком: рейтинг заемщика, подробная информация об имеющихся кредитах в других банках, детальная информация о просроченных или полностью погашенных прошлых кредитах, наличие других банковских продуктов и услуг у заемщика и пр.

Таким образом, независимые или скоринговые переменные могут быть представлены в разных шкалах измерения в зависимости от возможности объективных измерений интересующих признаков, используемых статистических методов для построения скоринговой модели. На практике могут быть построены скоринговые модели со следующими типами независимых переменных: только с количественными, только с категориальными[11], с категориальными и с количественными переменными.

2.4. Формирование обучающей и тестовой выборки

Доступные для построения скоринговой модели информационные данные называются часто исторической выборкой. Историческая выборка должна как можно точнее отражать исследуемую генеральную совокупность заемщиков, т.е. быть репрезентативной. Для проверки адекватности и точности предсказания скоринговой модели на этапе ее разработки историческую выборку необходимо разделить на две группы: обучающую выборку — наблюдения, по которым будет непосредственно строится модель; тестовую или контрольную выборку — наблюдения по которым будет известно значение зависимой переменной, но они не будут участвовать в построении модели, а будут использованы для проверки точности предсказания модели. Обучающая и контрольная выборка должна формироваться на основе механизма случайного отбора обычно в соотношении 70–80% и 30–20% соответственно от исходного объема исторической выборки.

Тестовая выборка используется после построения модели логистической регрессии для проверки ее достоверности[12]. Для кредитного скоринга — это прежде всего способность модели отличать «хороших» заемщиков от «плохих». Проверка достоверности модели заключается в ее применении и сравнении результатов на контрольной и тестовой выборке. Модель должна давать корректные прогнозы не только на обучающей совокупности, но и на практике при ее применении. Обычно используют стратегию генерализации модели на основе двух выборок. Схожие показатели точности, полученные на обучающей и тестовой выборке — признак того, что на практике скоринговая модель будет работать примерно также.

Более сложная стратегия генерализации модели предполагает формирование трех и более выборок: первая выборка используется для оценки параметров модели; вторая выборка используется для проверки модели, если получены значительные отклонения результатов по обучающей и тестовой выборке, то из них удаляются выбросы или переменные, влияющие на

[11] Количественные переменные в этом случае категоризируются на основе анализа зависимости их влияния на уровень дефолтов, что будет показано далее.

[12] Этот этап построения модели еще называют валидацией (от англ. validity — доказанность, обоснованность, пригодность). Тест модели на классификационную способность модели вне обучающей совокупности называют также генерализацией.

отклонения, и строится новая модель по объединенной первой и второй выборке; результаты новой модели тестируются на третьей выборке[13].

2.5. Определение объема выборки

Важным методологическим аспектом построения скоринговой модели является определение необходимого объема выборки. Обучающая выборка должна формироваться из исходной статистической базы на основе рассмотренных выше критериев случайным образом. Число принятых кредитных решений за заданный период времени должно быть достаточно большим для обеспечения необходимого объема выборки. Определение минимального объема выборки может опираться на следующие критерии: равномерность распределения зависимой переменной, число независимых переменных в модели, максимально допустимой ошибки выборки, экспертные оценки.

В литературе содержатся следующие рекомендации по определению минимального объема выборки и числа предикторов в модели[14]. Выбор минимального объема выборки зависит от равномерности распределения значений зависимой переменной. При относительно равномерном распределении необходимо задавать не менее 10 наблюдений на 1 предиктор. Но чем больше распределение зависимой бинарной переменной смещено в пользу в пользу одной из категорий, тем больше наблюдений нужно брать на один предиктор.

Другой подход, так называемое правило 20 EPV[15], также связывает минимальный объем выборки с распределением зависимой переменной и количеством предикторов в модели. Согласно этому подходу, необходимо взять количество наблюдений в исторической выборке, у которых зависимая переменная имеет наименьший объем, в кредитном скоринге, это «плохие» заемщики. Это число наблюдений нужно разделить на число предикторов, включенных в модель. На один предиктор должно приходиться не менее 20 наблюдений. Если это правило выполняется, то объем выборки достаточный.

В кредитном скоринге распределение зависимой переменной имеет всегда существенное смещение в пользу «хороших» заемщиков, поэтому рассмотренные критерии не подходят. Другой подход к определению объема выборки — на основе критерия мощности при задании максимально допустимой ошибки оценки соотношения «плохих» и «хороших» заемщиков в генеральной совокупности. Обучающая выборка заемщиков для построения модели должна отражать генеральную совокупность всех потенциальных заемщиков с неизвестными долями «плохих» и «хороших» кредитов. Предположим, мы хотим быть уверенным на 95%, что соотношение «плохих» и «хороших» заемщиков в обучающей выборке должно отражать генеральную популяцию заемщиков. Зная распределение зависимой переменной на тестовых данных, можно по формуле рассчитать необходимый объем выборки[16]:

[13] Существуют и другие стратегии валидации модели, например, стратегия прогнозирования постфактум, не предполагающая задание обучающей выборки, модель тестируется на реальных данных в течение определенного периода с последующей корректировкой ее параметров.

[14] См., например: (Hosmer, Lemeshow, 2000).

[15] См., например: (Harrell, Frank, 2001).

[16] Данная формула исходит из предположения формирования выборки простым случайным отбором из генеральной совокупности неизвестного объема. См., например: (Улитина и др., 2008, с. 263–267).

$$n = \frac{z_\gamma{}^2 w \cdot (1-w)}{\Delta_w^2} \quad (1),$$

где n — минимальный объем выборки, z_γ — значение стандартного нормального закона распределения, определяемое в зависимости от выбранного уровня надежности γ, w — доля «плохих» клиентов по тестовой выборке, Δ_w — максимально допустимая предельная ошибка оценки доли «плохих» заемщиков.

Например, среди 700 клиентов предварительной выборки 50 оказались «плохими». Оценка доли «плохих» клиентов по имеющимся данным для построения модели составила около 0.07 или 7%. При таком значении оценки доли, предположим, мы хотим ошибиться не более чем на 5%, что будет соответствовать допустимой предельной ошибке оценки доли 0.0035. При этом мы хотим получить результаты с надежностью не менее 99%. В этом случае z-значение стандартного нормального закона распределения составит около 2.58. Подставим эти значения в формулу (1) получим минимально необходимый объем выборки 35260. Такой объем выборки не всегда доступен. Если задать предельную ошибку выборки 10% или 0.007, а надежность 95%, что вполне приемлемо для построения предварительной модели на первом этапе, до ее калибровки, получаем совсем другие результаты: 5104 наблюдений.

При определении минимального объема выборки можно воспользоваться и формулой интервальной оценки генеральной доли:

$$w - z_\gamma \sqrt{\frac{w(1-w)}{n}} \le P \le w + t_\gamma \sqrt{\frac{w(1-w)}{n}}, \quad (2)$$

где P — оцениваемая доля «плохих» заемщиков в генеральной совокупности, w — доля «плохих» заемщиков по тестовой выборке, n — объем изначальной исторической выборки, z_γ — значение стандартного нормального закона распределения, определяемое в зависимости от выбранного уровня надежности γ.

По формуле (2) найдем верхнюю и нижнюю границу доверительного интервала для оценки генеральной доли «плохих» заемщиков для нашего примера при уровне надежности 95%: 0.05 и 0.09. В лучшем случае с точки зрения оценки риска, но наихудшем с точки зрения для определения минимального объема выборки, будет нижняя граница доверительного интервала для доли «плохих» клиентов. В нашем примере это около 0.05, и тогда необходимо будет взять объем выборки примерно 7299 клиентов при относительной ошибке оценки доли в 10%. А при доле «отрицательных» исходов в 0.09 будет уже достаточно 3885 наблюдений. Если же мы повысим точность оценивания генеральной доли до относительной ошибки в 5%, то уже понадобиться соответственно 29196 и 15537 наблюдений.

3. Анализ и корректировка переменных для построения модели

3.1. Корректировка распределения зависимой переменной

Одна из проблем, которая возникает после построения модели логистической регрессии в кредитном скоринге, связана с низкой степенью точности предсказания отрицательных исходов, т.е. дефолтов. Причина этого — недостаточное число «плохих» заемщиков в обучающей выборке. На практике обычно строят скоринговые карты на основе выборки, измеряемой тысячами наблюдений, с равным числом «положительных» и «отрицательных»

исходов зависимой переменой[17]. Зная истинные пропорции «плохих» и «хороших» заемщиков в генеральной совокупности и необходимый объем выборки, можно изменить распределение зависимой переменной с помощью методов перевзвешивания и прореживания выборки.

Первая стратегия подготовки данных для построения модели логистической регрессии при малом количестве «плохих» заемщиков и большом объеме выборки может быть следующей. Можно взять 100% всех «плохих» заемщиков и случайным образом отобрать часть «хороших» наблюдений. После реализации такого алгоритма доля «плохих» к «хорошим» наблюдениям может составлять от 0.1 до 0.5.

Вторая стратегия может быть более предпочтительна при небольшом объеме выборки. Суть ее в перевзвешивании данных, чтобы добиться нужного соотношения «плохих» и «хороших» наблюдений в выборке для построения модели. Перевзвешивание данных, в отличие от корректировки выборки за счет случайного отбора, дает, как правило, более надежные оценки параметров модели логистической регрессии. При таком подходе следует избегать двух основных ошибок, которые могут привести к смещению коэффициентов логистической регрессии. Во-первых, если брать часть «хороших» наблюдений, а не все, то их следует отбирать на основе механизма случайного отбора, чтобы наблюдения были независимыми. Например, отобрав каждое третье «хорошее» наблюдение в данных, мы рискуем получить зависимые наблюдения. Во-вторых, «плохие» и «хорошие» наблюдения должны быть сформированы по одной методике и на основе одинаковых критериев отбора данных, рассмотренных выше.

После корректировки выборки и оценки параметров модели логистической регрессии ее способность классифицировать редкие события улучшится. Коэффициенты же при независимых переменных останутся практически неизменными. Если логистическая регрессия используется для разработки скоринговой карты, то значений коэффициентов при независимых переменных будет достаточно. Меняя пропорции «успехов» и «неуспехов» в выборке при построении модели логистической регрессии, изменится оценка значения константы модели. Для оценки правильной вероятности наступления моделируемого события риска в случае перевзвешивания данных, значение константы корректируется[18].

В табл. 1 приведены результаты эмпирических расчетов коэффициентов логистической регрессии для условных переменных для иллюстрации рассмотренных методик формирования обучающей выборки. Изначально по выборке 700 наблюдений с долей «плохих» наблюдений 26% и «хороших» 74% точность предсказания «плохих» наблюдений составила всего 50%. Было реализовано два альтернативных подхода. В первом были взяты все «плохие» наблюдения и случайно отобрано 50% «хороших» наблюдений. Таким образом, соотношение «плохих» и «хороших» заемщиков составило 42% и 58% соответственно. Объем откорректированной выборки составил 434 наблюдений. Это позволило увеличить процент верно предсказанных «плохих» наблюдений до 71%. При этом доля верного предсказания «хороших» наблюдений составила 82%. Во втором подходе все данные были перевзвешены, чтобы соотношение «плохих» – «хороших» в выборке было также 42 и 58% процентов. Число наблюдений в выборке сохранилось и составило 700. Это привело практически к идентичным результатам по сравнению с первой методикой.

---

[17] Разработчики кредитных скоринговых систем под заказ, как правило, запрашивают выборку минимум из 4500 клиентов: 1500 «хороших», 1500 «плохих» и 1500 «отклоненных». Обычно объем выборки составляет около 2000 «плохих» и 2000 «хороших» заемщиков.

[18] Этот вопрос будет рассмотрен далее.

Таблица 1

Сравнение точности классификации при различных стратегиях формирования обучающей выборки

	Исходная модель	Отобраны все «плохие» и случайным образом 50% «хороших»	Исходные данные перевзвешены, чтобы увеличить долю плохих до 42%
Количество «плохих»	183	183	295
Доля плохих	0.26	0.42	0.42
Количество «хороших»	517	251	405
Доля «хороших»	0.74	0.58	0.58
Объем выборки	700	434	700
Коэффициенты модели			
Переменная 1	0.57	0.59	0.56
Переменная 2	– 0.08	– 0.07	– 0.10
Переменная 3	– 0.24	– 0.24	– 0.26
Переменная 4	0.09	0.09	0.09
Константа	– 0.79	– 0.22	– 0.95
Классификация			
Процент корректных «хороших»	92.5	82.1	79.9
Процент корректных «плохих»	50.3	70.5	73.8
Общий процент корректных	81.4	77.2	77.3

3.2. Описательный анализ скоринговых переменных

Перед построением модели логистической регрессии данные следует подвергнуть подробному и всестороннему статистическому анализу. Исследование данных решает две главные задачи: 1) позволяет обнаружить возможные ошибки и пропуски в данных; 2) позволяет оценить силу связи между независимыми переменными и зависимой. Для категориальных скоринговых переменных полезным будет вывод частотных таблиц распределений. Категории независимых переменных с малым числом наблюдений, как правило, при возможности объединяют с соседними[19] категориями. Также частотные таблицы позволяют обнаружить недопустимые и ошибочно введенные значения в данных.

Для количественных переменных такие базовые описательные статистики как среднее значение, медиана, стандартное отклонение, процентные точки, доля пропущенных значений может дать представление о целостности и состоятельности данных. Визуальное представление

[19] Число наблюдений в каждой категории зависимой переменной должно быть, по разным оценкам, не менее 3–5% от общего числа валидных значений по данной категориальной переменной.

данных дают с помощью гистограмм, «ящичковых» диаграмм. Распределения скоринговых переменных по выборке необходимо сравнить с распределениями во всем кредитном портфеле для проверки репрезентативности обучающей выборки. Отдельное внимание уделяется наличию в данных допустимых значений, но имеющих экстремальный характер. При незначительном количестве таких величин их, как правило, удаляют из анализа или заменяют средними значениями, предварительно исследовав, не являются ли они признаками мошеннических кредитов.

Работа с пропущенными значениями — отдельный этап анализа данных перед построение модели. Огромные массивы финансовых данных для построения скоринговой модели всегда содержат на практике пропуски, причину которых необходимо проанализировать. Причиной пропусков могут быть: ошибки при сборе и вводе данных, невозможность получения информации, сознательный отказ от ответа заемщиком. Наблюдения с пропусками в переменных по причине ошибок ввода данных следует исключить из анализа или заменить средними значениями переменных.

Построение модели логистической регрессии требует полного набора данных и одинакового числа наблюдений по каждой переменной. При наличии пропусков в данных число наблюдений для каждой переменной будет различаться. Обычно считается, что наличие пропусков в менее 5% наблюдений можно объяснить случайностью[20]. В этом случае наблюдения с пропусками можно просто удалить, исключив таким образом их из анализа без существенной потери информации и ухудшения качества статистической модели. Наличие в данных наблюдений с более чем 5% пропусков может быть не случайно, а закономерно. Их удаление может привести к искажению результатов моделирования в связи с существенным сокращением количества наблюдений в выборке. В этом случае необходимо выявить причину пропусков и восстановить пропущенные значения, либо заменить их. Наблюдения или переменные, которые имеют значительное число пропущенных значений (более 50%) обычно исключают из анализа.

В случае достаточного количества пропусков особенно по причине осознанного отказа от ответа заемщиками следует пропущенные значение перекодировать в отдельную категорию и включить их в анализ, наравне с другими категориями[21]. Такой подход предполагает неслучайный характер пропусков, в ряде случаев пропуски в данных являются дополнительным индикатором «плохого» заемщика. Например, если заемщик имеет низкий доход, то он с большой вероятностью оставить пустым поле «Ваш доход» в анкете-заявке. Отдельное направление работы с пропущенными данными — их импутация на основе скрытых закономерностях в данных. В основе импутации данных лежат алгоритмы получения возможных значений для отсутствующих данных по значениям других валидных переменных на основе многомерных статистических методов[22].

3.3. Преобразование количественных переменных

Моделирование вероятности дефолта может включать этап поиска непрерывных преобразований для скоринговых количественных характеристик. Независимые количественные переменные могут включаться в модель логистической регрессии без

[20] Существуют специальные статистические тесты для оценки случайности пропусков, см. например: (Little, 1988).

[21] Иногда эти категории вводят с весами, близкими к медианному или среднему значению переменной, в которой анализируются пропуски.

[22] См., например: (Rubin, 1987), (Schafer, 1997).

преобразований и на основе непрерывных преобразований. Обычно для преобразования распределения количественных переменных используют следующие виды преобразований: квадратное; кубическое; квадратный корень; натуральный или десятичный логарифм; экспоненциальное; величина, обратная квадратному корню; обратная величина. При использовании степенных преобразований ко всем значениям преобразуемой переменной могут добавлять константу для преобразования нуля или отрицательных значений. Такие преобразования количественных переменных могут привести к максимизации их связи с зависимой целевой переменной. Нужный вид преобразований можно определить на данных конкретной выборки эмпирическим путем[23]. Лучшее преобразование имеет наибольшую значимую тесноту связи с зависимой переменной. Окончательный выбор наилучшего варианта преобразования происходит на основе сравнения характеристик точности полученных уравнений логистической регрессии.

При рассмотрении количественных независимых переменных в модель также часто вводят их относительные преобразования. Примерами таких расчетных переменных могут быть отношение ежемесячных выплат по кредиту к среднему ежемесячному доходу заемщика, отношение месячного свободно располагаемого бюджета заемщика к ежемесячному доходу, отношение суммы задолженности к доходу и др.

3.4. Оценка мультиколлинеарности между количественными переменными

При включении в модель логистической регрессии количественных переменных необходимо проанализировать их на наличие мультиколлинеарности. Этой проблеме построения модели логистической регрессии уделяются недостаточно внимания, хотя она также актуальна при использовании количественных предикторов, как и для модели линейной множественной регрессии. Первоначальный анализ наличия мультиколлинеарности может быть произведен на основе матрицы парных и частных корреляций между независимыми переменными. Однако коэффициенты корреляции не всегда могут показать наличие мультиколлинеарности. Для диагностики мультиколинеарности часто используют показатели толерантности или допуска переменной, определяемой по формуле:

$$1 - R_i^2, \tag{3}$$

где R_i^2 — квадрат множественного коэффициента корреляции *i*-ой независимой переменной со всеми остальными предикторами. Если толерантность переменной близка к 0, то значения данной переменной можно выразить через линейную комбинацию остальных независимых переменных.

Иногда вместо толерантности используют показатель, обратный ее величине, называемый коэффициентом или фактором «вздутия» дисперсии:[24]

$$VIF = \frac{1}{1 - R_i^2}. \tag{4}$$

При наличии мультиколлинеарности дисперсия оценок параметров модели регрессии возрастает пропорционально данной величине, что делает их оценку нестабильной. Большое значение показателя VIF свидетельствует о наличии мультиколлинеарности. Общепринятое

[23] Конкретные виды таких преобразований зачастую составляют коммерческую тайну готовых скоринговых моделей и не разглашаются.

[24] От англ. Variance Inflation Factor или VIF.

мнение, что если этот показатель больше 5, это свидетельствует о наличии мультиколлинеарности. Другое мнение, что этот показатель должен быть больше 10[25].

В случае наличия мультиколлинеарности необходимо найти оптимальные варианты исключения тесно коррелирующих переменных для построения модели. Другой метод борьбы с мультиколлинеарностью — это увеличение объема выборки. Следует отметить, что на больших выборках объемом в несколько тысяч и десятков тысяч заемщиков мультиколлинеарность не сильно искажает оценки параметров модели. Оценки параметров на больших выборках получаются статистически устойчивыми. Часто помогающий способ избавления от мультиколлинеарности — включение в модель нелинейных преобразований от тесно связанных независимых переменных. Кардинальный метод избавления от мультиколлинеарности — использование главных компонент вместо исходных значений переменных — используется при построении скоринговых моделей редко из-за последующей сложности расчета скоринговых баллов на основе агрегированных показателей.

3.5. Категоризация количественных переменных

Как уже отмечалось выше, в скоринговой модели могут использоваться в качестве независимых переменных категориальные и количественные предикторы. Единого мнения по поводу выбора наилучшей модели для эффективных скоринговых карт не существует. Многие разработчики скоринговых систем используют всегда метод категоризации количественных переменных. Другие разработчики находят для отражения нелинейной зависимости вероятности дефолта непрерывные преобразования[26]. Но исторически сложилось, что чаще для построения скоринговых карт используют категориальные предикторы. Категоризация количественных переменных позволят добиться следующих основных преимуществ при построении скоринговой карты: облегчить обработку выбросов и экстремальных значений количественных переменных; упростить интерпретацию скоринговой карты; отразить сложные нелинейные связи.

Категоризация количественных переменных происходит по следующему алгоритму[27]. Первоначально количественная переменная разбивается на основе равных процентилей на несколько групп[28]. Затем в каждой группе считается доля «плохих» и «хороших» кредитов, а также показатель веса категорий предиктора WOE[29]. Веса категорий предиктора помогают найти по переменной «границы чувствительности» к появлению моделируемого события риска и провести оптимальным образом категоризацию количественных переменных. Показатели WOE для каждой категории рассчитываются по формуле:

$$WOE_i = \ln\left(\frac{d_i^{(1)}}{d_i^{(2)}}\right), \quad (5)$$

[25] Существуют и другие менее распространенные методы анализа мультиколлинеарности: анализ собственных значений, получастных или частичных коэффициентов корреляций, показателей обусловленности, определителя корреляционной матрицы и др.

[26] Для решения этой проблемы компания Experian провела в 2000 году исследование. Часть результатов этого исследования и сравнительный анализ моделей с категориальными и непрерывными переменными приведены в (Mays et al., 2001).

[27] Этот алгоритм может еще называться процедурой биннинга.

[28] Может использоваться разбиение на основе от 10 до 50, а иногда и более процентилей.

[29] От англ. Weight of Evidence.

где $d_i^{(1)}$ и $d_i^{(2)}$ — относительные частоты «плохих» и «хороших» кредитов соответственно в *i*-ой группе категоризованной переменной; *i*=1, 2, … , *k*; *k* — число категорий переменной.

Далее полученные показатели весов категорий анализируются, происходит объединение соседних категорий и расчет показателей WOE повторяется. Пропущенные данные при категоризации могут кодироваться как отдельная категория и участвовать в дальнейшем анализе. При дальнейшем объединении категорий руководствуются следующими правилами: в каждой группе должно находиться не меньше 5% от всех валидных наблюдений переменной; не должно быть групп с количеством «плохих» или «хороших» кредитов, равным 0; процент «плохих» заемщиков и WOE должны в достаточной мере отличаться по получаемым группам; значения показателей WOE должны иметь возрастающий или убывающий тренд при переходе от одной категории к другой. При укрупнении категорий помимо статистических критериев следует руководствоваться логикой, целесообразностью и возможностью такого объединения.

Проиллюстрируем рассмотренную методику биннинга на примере количественной переменной стажа работы при построении одной и скоринговых карт. В табл. 2 приведены рассчитанные показатели веса категорий (WOE).

Таблица 2

Рассчитанные показатели WOE на основе децилей

Группы переменной стаж работы на основе децилей		Наличие дефолта по кредиту		Итого	% не дефолтов	% дефолтов	WOE	IV
		Нет	Да					
Стаж работы на последнем месте, лет	<= 1	60	51	111	0.12	0.28	0.88	0.14
	2	19	25	44	0.04	0.14	1.31	0.13
	3 – 4	58	31	89	0.11	0.17	0.41	0.02
	5 – 6	61	21	82	0.12	0.11	– 0.03	0.00
	7 – 7	30	8	38	0.06	0.04	– 0.28	0.00
	8 – 9	62	14	76	0.12	0.08	– 0.45	0.02
	10 – 12	48	8	56	0.09	0.04	– 0.75	0.04
	13 – 14	57	14	71	0.11	0.08	– 0.37	0.01
	15 – 18	66	7	73	0.13	0.04	– 1.21	0.11
	19+	56	4	60	0.11	0.02	– 1.60	0.14
Итого		517	183	700	1.00	1.00		0.62

Проанализируем значения WOE в табл. 2 и объединим категории с близкими значениями WOE в одну категорию[30], получив новую переменную стажа работы с 5 категориями. Для новой

[30] В таблице объединяемые строки выделены серым цветом.

переменной повторно рассчитаем значения WOE (см. табл. 3). На рис. 1 видно, что благодаря процедуре биннинга удалось получить новую переменную с меньшим числом категорий, для каждой из которых значение WOE уменьшается с ростом значения категории.

Таблица 3

Рассчитанные показатели WOE после укрупнения категорий

Группы переменной стаж работы на основе 5 категорий		Наличие дефолта по кредиту		Итого	% не дефолтов	% дефолтов	WOE	IV
		Нет	Да					
Стаж работы на последнем месте, лет	<= 2	79	76	155	0.15	0.42	1.00	0,26
	3 – 4	58	31	89	0.11	0.17	0.41	0,02
	5 – 6	25	11	36	0.05	0.06	0.22	0,00
	7 – 14	233	54	287	0.45	0.30	– 0.42	0,07
	15 +	122	11	133	0.24	0.06	– 1.37	0,24
Итого		517	183	700	1,00	1.00		0.59

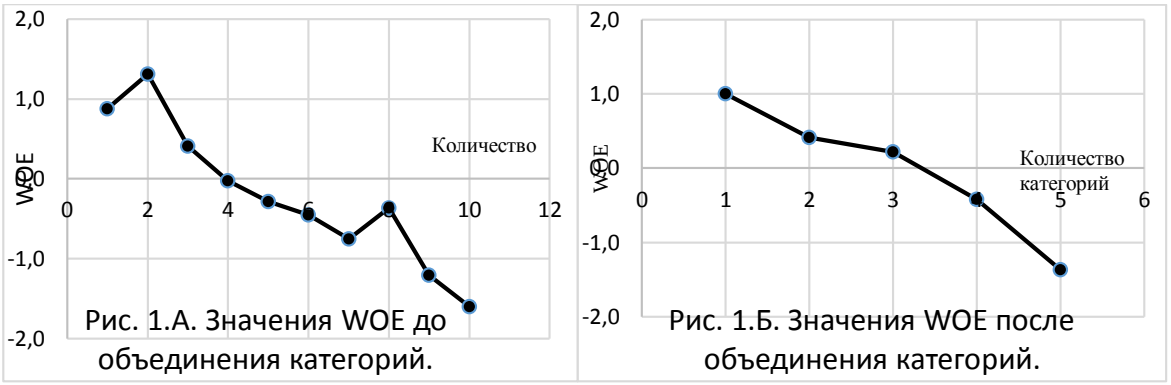

Рис. 1. *Изменение предсказательной способности категорий в зависимости от их числа*

3.6. Оценка взаимосвязи скоринговых переменных на вероятность дефолта

Предварительный анализ взаимосвязи скоринговых переменных на вероятность дефолта по кредиту помогает ограничить количество рассматриваемых для построения модели логистической регрессии переменных. Основными методами оценки наличия связи между зависимой бинарной переменой и независимыми категориальными переменными является расчет критерия хи-квадрат и показателя информационного значения[31].

При использовании критерия хи-квадрат выдвигают гипотезу H_0 об одинаковом распределении «плохих» и «хороших» заемщиков по категориям независимой переменной. По

[31] Как правило, скоринговые карты строятся по категориальным переменным, поэтому ограничимся рассмотрением модели только с категориальными независимыми переменными. Количественные независимые переменные должны быть категоризованы с помощью биннинга.

таблице сопряженности между зависимой и каждой анализируемой отдельно независимой переменной рассчитывают статистику χ^2-критерия по формуле:

$$\chi^2_P = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(f_{ij}-e_{ij}\right)^2}{e_{ij}}, \qquad (6)$$

где f_{ij} – фактические частоты; e_{ij} – ожидаемые частоты; m и k – число строк и столбцов в таблице сопряженности.

Альтернативой формулы (6) является расчет критерия хи-квадрат на основе формулы логарифма правдоподобия:

$$\chi^2_{LR} = 2\sum_{i=1}^{m}\sum_{j=1}^{k} f_{ij}\ln\left(\frac{f_{ij}}{e_{ij}}\right). \qquad (7)$$

При достаточном числе наблюдений значение по альтернативной формуле (7) будет мало отличаться от значения по классической формуле (6). Если расчетное значение критерия будет превышать критическое значение по таблице распределения хи-квадрат с заданным уровнем значимости и числом степеней свободы $\chi^2_{кр} = \chi^2\left(\alpha;\nu=(m-1)(k-1)\right)$, то проверяемая гипотеза будет отвергаться, будет доказано наличие взаимосвязи между анализируемой независимой переменной и вероятностью дефолта по кредиту.

Таблица 4

Значения частот для расчета критерия хи-квадрат

Наличие дефолта по кредиту		Стаж работы, лет					Итого
		<2	3–4	5–6	7–14	15>	
Нет	Частота	79	58	25	233	122	517
	Ожидаемая частота	114.5	65.7	26.6	212.0	98.2	517.0
Да	Частота	76	31	11	54	11	183
	Ожидаемая частота	40.5	23.3	9.4	75.0	34.8	183.0
	Частота	155	89	36	287	133	700
	Ожидаемая частота	155.0	89.0	36.0	287.0	133.0	700.0

В табл.4 приведены фактические и ожидаемые частоты для рассматриваемого выше примера зависимости дефолта по кредиту и категорий стажа работы. По данной таблице значение статистики хи-квадрат по формулам (6) и (7) будут равны соответственно 75.88 и 76.73 Критическое значение при уровне значимости менее 0.001[32] и 4 степенях свободы будет меньше рассчитанных значений, наличие взаимосвязи между стажем работы и наличием дефолта по кредиту доказана.

В качестве меры связи между независимыми переменными и зависимой часто используют коэффициент Крамера, коэффициент сопряженности Пирсона, коэффициент Фишера. Однако данные коэффициенты рассчитываются на основе значения статистики хи-

[32] Статистические пакеты, рассчитывающие значение теста хи-квадрат обычно выдают наименьший уровень значимости, при котором можно отвергнуть проверяемую гипотезу, а не уровень значимости задаваемый исследователем.

квадрат и являются фактически переводом расчетного значения хи-квадрат в шкалу от 0 до 1. Поэтому эти коэффициенты говорят о степени значимости результатов теста хи-квадрат и оценивают только наличие взаимосвязи, не показывая причинно-следственную связь. Для оценки степени взаимосвязи между независимыми переменными и бинарной зависимой в кредитном скоринге принято использовать расчет показателя информационного значения или IV[33] по формуле:

$$IV = \sum_{i=1}^{k} (d_i^{(1)} - d_i^{(2)}) \cdot WOE_i \,, \qquad (8)$$

где k — число категорий независимой переменной, остальные обозначения из формулы (5).

Для нашего примера расчет значений IV приведен в последней графе табл. 2 и табл. 3. Чем выше информационное значение предиктора, тем больший вес он имеет с точки зрения полезности при построении модели. Можно руководствоваться следующими правилами при отборе переменных для построения модели логистической регрессии: если значение IV менее 0.02 — независимая переменная не обладает прогностической способностью; от 0.02 до 0.1 — низкая прогностическая способность; от 0.1 до 0.3 — средняя прогностическая способность; от 0.3 до 0.5 — высокая прогностическая способность; более 0.5 — превосходная прогностическая способность.

4. Построение скоринговой карты

4.1. Модель логистической регрессии

Логистическая регрессия — самая распространенная статистическая модель для построения скоринговых карт при бинарной зависимой переменной. Математически модель логистической регрессии выражает зависимость логарифма шанса (логита) от линейной комбинации независимых переменных:

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_i^{(1)} + b_2 x_i^{(2)} + \ldots + b_k x_i^{(k)} + \varepsilon_i \,, \qquad (8)$$

где p_i — вероятность наступления дефолта по кредиту для i-го заемщика; $x_i^{(j)}$ — значение j-ой независимой переменной; b_0 — независимая константа модели, b_j — параметры модели; ε_i — компонент случайной ошибки.

Уравнение (8) отражает линейную зависимость вероятности наступления просрочки по кредиту в зависимости от значений независимых переменных. Константа в модели отражает естественный уровень риска наступления моделируемого события при равенстве всех независимых переменных нулю. Значения коэффициентов при независимых переменных, отражающих степень их влияния на шанс дефолта в логарифмической шкале, используются для построения скоринговой карты.

Значение константы в модели логистической регрессии зависит от распределения в данных по категориям зависимой переменной. В случае перевзвешивания выборки для изменения этого распределения для более адекватной последующей оценки качества

[33] От англ. IV — Information Value.

полученной модели константу корректируют и получают следующую модель логистической регрессии:

$$\ln\left(\frac{p_i^*}{1-p_i^*}\right)=\ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right)+b_0+b_1x_i^{(1)}+b_2x_i^{(2)}+\ldots+b_kx_i^{(k)}+\varepsilon_i\,, \qquad (9)$$

где p_i^* — откорректированная априорная вероятность; ρ_0 и ρ_1 — доли «хороших» и «плохих» заемщиков в выборке; π_0 и π_1 — доли «хороших» и «плохих» заемщиков в генеральной совокупности.

Исходя из модели (8) могут быть откорректированы значения прогнозируемых апостериорных вероятностей дефолта и получены, таким образом априорные вероятности для генеральной совокупности[34]. Однако для построения скоринговой карты достаточно значений коэффициентов при независимых переменных, которые при преобразовании (9) остаются неизменными.

Для интерпретации коэффициентов модели логистической регрессии обычно используют экспоненциальную форму записи модели:

$$p_i=\frac{1}{1+\exp(-(b_0+b_1x_i^{(1)}+b_2x_i^{(2)}+\ldots+b_kx_i^{(k)}+\varepsilon_i))}\,. \qquad (10)$$

При включении в модель логистической регрессии непрерывных количественных переменных коэффициенты при них будут показывать, на сколько в среднем измениться логарифм шанса наступления просрочки по кредиту при изменении независимой переменной на единицу своего измерения при неизменности остальных переменных. В экспоненциальной форме коэффициенты будут показывать насколько в среднем изменятся шансы наступления дефолта при изменении независимой переменной на единицу своего измерения при неизменности остальных переменных. Если коэффициент регрессии будет положительный, то его экспонента будет больше единицы и шансы будут возрастать, если коэффициент окажется отрицательным — меньше, шансы будут убывать. При включении в модель бинарной независимой переменной, коэффициент регрессии в экспоненциальной форме при фиктивной переменной будет показывать соотношение шансов проявления дефолтов при наличии фактора, отражаемого бинарной независимой переменной, по сравнению с его отсутствием[35].

4.2. Методы и способы включения независимых переменных в модель

Остановимся на основных способах включения в модель логистической регрессии предварительно отобранных независимых переменных, имеющих значимую связь на зависимую. Первый метод предполагает включение в модель количественных предикторов без их категоризации. Для устранения различий в единицах измерения количественных переменных и сравнения коэффициентов регрессии между собой исходные количественные переменные часто стандартизируют. В случае наличия категориальных независимых переменных они могут вводится в модель с помощью нескольких бинарных переменных. На каждую категорию создается своя бинарная переменная, принимающая значения 0 и 1, при этом одна из категорий берется за эталон сравнения и для нее не создается категориальная переменная. Это стандартный подход построения модели логистической регрессии.

[34] См. подробнее: (Naeem, 2009, p. 68).

[35] Такие коэффициенты еще называют коэффициентами соотношения шансов.

В кредитном скоринге чаще используют второй подход включения переменных в модель. Для построения модели используются только категориальные переменные. Если имеются количественные переменные их предварительно категоризуют с помощью процедуры биннинга. Затем для всех без исключения градаций категориальных переменных рассчитывают показатели WOE и заменяют ими фактические значения независимых переменных. По значениям WOE строят модель логистической регрессии.

В итоговой скоринговой карте должно быть достаточное число атрибутов для ее стабильной работы[36]. При отборе переменных учитывают такие моменты, как: независимость между объясняющими переменными, итоговую прогностическую способность модели, интерпретируемость получаемых коэффициентов при значимых переменных. Для построения окончательной модели логистической регрессии и отбора финальных переменных используют стандартные пошаговые алгоритмы включения, исключения переменных или их сочетание для получения значимого уравнения регрессии с достаточной прогностической способностью и со всеми значимыми коэффициентами регрессии.

4.3. Критерии качества модели логистической регрессии

Математическим аспектам построения модели логистической регрессии, оценке ее параметров в литературе уделено достаточно внимания, поэтому остановимся на общем алгоритме построения уравнения логистической регрессии и обзоре основных критериев ее качества[37].

Оценки параметров модели логистической регрессии находят методом максимального правдоподобия. Общей оценкой качества подгонки модели логистической регрессии может служить значение функции правдоподобия. На практике значение функции правдоподобия преобразуют через минус удвоенное значение логарифма правдоподобия, поскольку такое преобразование имеет распределение хи-квадрат, с помощью которого проверяется гипотеза о значимости модели в целом. Для хорошей модели функция правдоподобия близка к 1, а минус удвоенный логарифм функции правдоподобия близок к 0.

Дополнительным тестом для оценки качества подгонки модели является тест Хосмера–Лемешева. Этот тест позволяет проверить гипотезу о соответствии наблюдаемых и спрогнозированных значений зависимой переменной и является альтернативной величиной качества модели. На результаты этого теста следует особо обращать внимание при включении в модель предикторов, измеренных в непрерывной шкале, и выборках с небольшим числом наблюдений. Общий алгоритм расчета критерия Хосмера–Лемешева следующий: по расчетным значениям вероятностей зависимой переменной рассчитывают децили — делящие значение предсказанной вероятности дефолта на 10 групп[38]; строят таблицу сопряженности строки, которой задают группы децилей риска, а столбцы зависимая бинарная переменная; на основе критерия согласия хи-квадрат сравнивают степень различий фактических и ожидаемых частот в полученной таблице сопряженности.

После проверки значимости уравнения логистической регрессии проверяется значимость отдельных коэффициентов. При значимом уравнении регрессии будет по крайней мере один предиктор, объясняющий изменение зависимой переменной. В итоге желательно

[36] В идеале получить модель логистической регрессии с 8–15 независимыми переменными, значимо влияющими на дефолт.

[37] См. например: (Hosmer, Lemeshow, 2000).

[38] Эти группы часто называют децилями риска.

получить значимое уравнение со всеми значимыми коэффициентами. Если коэффициент при каком-то предикторе незначим, то его, как правило, исключают из состава независимых переменных и заново пересчитывают уравнение. При этом можно использовать алгоритмы пошагового отбора переменных для построения модели. Для проверки гипотезы о значимости отдельных коэффициентов используют статистику Вальда, имеющую распределение хи-квадрат. При проверке значимости коэффициентов необходимо обращать внимание и на абсолютные значения стандартных ошибок коэффициентов и доверительные интервалы. Чем меньше стандартная ошибка и уже доверительный интервал у коэффициентов в уравнении — тем предпочтительнее модель.

Для оценки качества подгонки модели часто используют значение коэффициента детерминации. Однако для модели логистической регрессии коэффициент детерминации не является основной характеристикой точности модели в отличие от модели линейной регрессии[39]. Коэффициенты детерминации в логистической регрессии используют для оценки меры зависимости между переменными на этапе построения модели и отбора предикторов. Псевдо коэффициенты детерминации не следует рассматривать как главные меры качества модели. Псевдо коэффициенты детерминации смешивают силу эффекта с качеством подгонки модели. Особенно это проявляется на малых выборках, когда значение коэффициента детерминации может быть высоким при неудовлетворительном качестве подгонки. Низкие значения коэффициентов детерминации в модели логистической регрессии — это нормальное явление. Оценка коэффициента детерминации в модели логистической регрессии может строится на основе логарифма функции правдоподобия, через сравнение фактических значений зависимой переменной и расчетных значений вероятностей зависимой переменной и др. способами.

Важной характеристикой любой регрессионной модели является достоверность. Достоверность модели логистической регрессии характеризуется ее способностью отличать «хороших» заемщиков от «плохих». Дискриминирующую способность модели можно оценить, проанализировав таблицу классификации. Важно построить модель логистической регрессии, одинаково хорошо различающей и «хороших» и «плохих» заемщиков. Для оценки качества классификации модели строят ROC–кривую[40], которая показывает зависимость количества верно классифицированных положительных исходов от количества неверно классифицированных отрицательных исходов. Для сравнения двух и более моделей между собой сравниваются площади под ROC–кривыми — этот показатель называется AUC[41] и измеряется от 0.5 до 1.

Чем больше значение площади, тем лучше модель. Обычно считают, что значение площади от 0.9 до 1 соответствует отличному качеству модели, от 0.8–0.9 — очень хорошему, 0.7–0.8 — хорошему, 0.6–0.7 — среднему, 0.5–0.6 — неудовлетворительному. По значению площади под ROC–кривой можно вычислить показатель индекс Джинни. Этот показатель переводит значение площади под кривой в диапазон от 0 до 1.

Скоринговые карты разрабатываются для ранжирования заемщиков по шансам наступления просрочки по кредиту. Важно, чтобы в скоринговой карте кредиты, по которым

---

[39] В отличие от линейной регрессии в логистической регрессии нельзя выдвинуть предположение о постоянстве дисперсии, поскольку дисперсия бинарной переменной зависит от частоты распределения значений самой переменной. Поэтому вычисляемые коэффициенты детерминации являются приближенной мерой и называются еще псевдо коэффициентами детерминации.

[40] От англ. Receiver Operator Characteristic.

[41] От англ. Area Under Curve.

происходит данное событие, и кредиты, с которыми оно не происходит, имели разные баллы. Чем более явно разделены распределения скоринговых баллов для «плохих» и «хороших» кредитов, тем эффективнее будет работать скоринговая карта. Для оценки качества построенной по модели логистической регрессии скоринговой карты также анализируют распределение скоринговых баллов по «плохим» и «хорошим» заемщикам. Для оценки качества прогнозирования модели логистической регрессии на основе этого распределения рассчитывают тест Колмогорова–Смирнова. В тесте Колмогорова–Смирнова сравниваются два кумулятивных распределения скоринговых баллов «хороших» и «плохих» заемщиков. Статистика Колмогорова–Смирнова вычисляется как максимальная разница между кумулятивными функциями этих распределений. Диапазон значений статистики Колмогорова–Смирнова измеряется от 0 до 100. Чем выше значение статистики Колмогорова–Смирнова, тем лучше работает модель. Альтернативной мерой оценки валидации модели логистической регрессии может быть коэффициент дивергенции, который представляет собой оценку разницы математических ожиданий распределений скоринговых баллов для «плохих» и «хороших» заемщиков, нормализованную дисперсиями этих распределений. Чем больше значение коэффициента дивергенции, тем лучше качество модели с точки зрения ее классификационной способности.

4.4. Перевод коэффициентов модели в скоринговую карту

Заключительным этапом разработки скоринговой модели является перевод коэффициентов логистической регрессии в скоринговые баллы. Если взять оценки коэффициентов логистической регрессии и умножить их на значения независимых переменных, то получиться итоговый скоринговый балл в шкале натуральных логарифмов:

$$итоговый\ \ балл = \hat{b}_1 x_1 + \hat{b}_2 x_2 + \ldots + \hat{b}_k x_k , \qquad (11)$$

где x_j — значение предикторов для оцениваемого заемщика, $\hat{b}_j$ — оценки коэффициентов логистической регрессии.

Для приведения скоринговых баллов в линейную шкалу используют прием масштабирования. Масштабирование не изменяет прогностическую способность скоринговой карты, а лишь переводит скоринговые баллы в новую шкалу, удобную для использования[42]. Скоринговый балл в линейной шкале представляет собой отношение шансов «хороших» заемщиков к «плохим». Для масштабирования необходимо прежде всего задать диапазон числовой шкалы с минимум и максимум (например, от 0 до 1000). На результат масштабирования также влияют два показателя: количество баллов, которое удваивает шансы стать «хорошим» заемщиком и значение шкалы, в котором достигается заданное отношение шансов «хороших» к «плохим». Наиболее часто используют скоринговые карты, в которых каждые 20 баллов удваивают шансы стать «хорошим». Другой стандарт — каждые 40 баллов удваивают шансы стать «хорошим» заемщиком. Для приведения коэффициента логистической регрессии в скоринговый балл в линейной шкале применяют следующее преобразование:

$$балл = A + R \cdot \hat{b}_j , \qquad (12)$$

где R — множитель; A — смещение.

[42] Часто на практике для различных скоринговых карт в банке используют непересекающиеся по значениям баллов друг с другом шкалы во избежание путаницы при назначении баллов заемщику.

Множитель определяют по формуле:

$$R = \frac{D}{\ln(2)}, \quad (13)$$

где D — количество баллов, удваивающее шансы.

Смещение определяют по формуле:

$$A = B - R \cdot \ln(C), \quad (14)$$

где B — значение на шкале баллов, в которой соотношение шансов составляет C:1.

Приведем пример расчета скоринговых баллов для одного атрибута скоринговой карты — стажа работы для рассмотренного выше примера. В табл. 5 приведены результаты расчетов параметров логистической регрессии[43]. Все параметры получились значимыми. Категориальная переменная была введена в модель с помощью нескольких бинарных переменных. За опорную категорию была выбрана последняя категория с максимальным стажем работы.

Таблица 5

Результаты расчетов коэффициентов логистической регрессии

Категории независимой переменной	Коэффициент регрессии B	Стандартная ошибка	Статистика Вальда	Значимость	Exp(B)
<= 2	2.367	0.353	44.865	0.000	10.670
3 – 4	1.780	0.385	21.313	0.000	5.928
5 – 6	1.585	0.480	10.924	0.001	4.880
7 – 14	0.944	0.349	7.311	0.007	2.570
Константа	-2.406	0.315	58.417	0.000	0.090

Будем предполагать, что каждые 40 баллов удваивают шансы наступления дефолта по кредиту, а в точке 600 баллов отношение шансов составляет 72:1[44]. По формуле (13) множитель в этом случае будет равен 57.71, а смещение по формуле (14) будет равно 413.20. Поскольку с увеличением стажа работы шансы дефолта по кредиту падают, множитель нужно брать с отрицательным знаком, чтобы значения скоринговых баллов возрастали с ростом стажа. В табл.6 приведен расчет скоринговых баллов по формуле (12). В моделях с несколькими независимыми переменными для получения общего скорингового балла необходимо сложить баллы по каждой независимой переменной.

Таблица 6

Расчет скоринговых баллов

Категории независимой переменной, лет	Коэффициент регрессии b_j	Скоринговый балл в линейной шкале b_j *R	Скоринговый балл с учетом смещения A – b_j *R
<= 2	2.367	136.6	550
3 – 4	1.780	102.7	516
5 – 6	1.585	91.48	505

[43] Расчеты осуществлялись в программе IBM SPSS Statistics 22.

[44] Такие предположения являются одним из общепринятых стандартов расчета скоринговых баллов.

7 – 14	0.944	54.5	468
15+	0	0	413

Если уравнение логистической регрессии строиться по значениям WOE, то формула расчета скорингового балла в линейном масштабе будет следующая:

$$балл = -\left(WOE_j \cdot b_i + \frac{b_0}{n}\right) \cdot R + \frac{A}{n}, \quad (15)$$

где WOE_j — значение WOE для каждой j-ой категории сгруппированной переменной, n — количество независимых переменных в уравнении регрессии, b_0 — константа, b_i — коэффициент регрессии для i-ой переменной.

В табл. 7 приведены результаты расчета модели логистической регрессии с одной независимой переменной категорий стажа работы рассматриваемого примера после замены категорий их весами WOE, а в табл. 8 расчет скоринговых баллов на основе этой модели. Отметим, что расчет скоринговых баллов по этой методике во многом зависит от того, насколько правильно была проведена категоризация количественных переменных, насколько существенна разница WOE между категориями.

Таблица 7

Результаты расчетов коэффициентов логистической регрессии по значениям WOE

Переменные	Коэффициент регрессии B	Стандартная ошибка	Статистика Вальда	Значимость	Exp(B)
Стаж работы	1.000	0.122	67.031	0.000	2.718
Константа	– 1.039	0.093	125.814	0.000	0.354

Таблица 8

Расчет скоринговых баллов при построении регрессии по WOE

Категории независимой переменной, лет	WOE_j	Скоринговый балл
<= 2	1.00	274
3 – 4	0.41	329
5 – 6	0.22	366
7 – 14	– 0.42	377
15+	– 1.37	411

Заключение

Один из основных инструментов снижения рисков — использование автоматизированных систем скоринга. В основе работы скоринговых систем лежит автоматический расчет баллов в зависимости от параметров запрашиваемого кредита, кредитной истории, социально-демографических характеристик заемщика. В зависимости от количества набранных баллов скоринговая система выдает решение: выдавать или не выдавать кредит клиенту банка. Большинство скоринговых систем строится на основе модели логистической регрессии. Коэффициенты полученного уравнения логистической регрессии масштабируются в скоринговые баллы. В работе рассматривалась методика построения скоринговых карт на базе модели логистической регрессии. Были изложены методические подходы к формированию и исследованию характеристик заемщика для построения модели,

формированию списка переменных для моделирования, рассмотрены возможные преобразования данных для успешного построения модели. Отдельное внимание в работе уделено исследованию качества и прогностических свойств модели логистической регрессии. На примере изложена техника перевода коэффициентов логистической регрессии в скоринговые баллы. Распространенной практикой является нахождение коэффициентов логистической регрессии по значениям весов категориальных переменных с последующим их переводом в скоринговые баллы.

ЛИТЕРАТУРА

1. Сорокин, А.С. Применение законов распределения случайных величин для моделирования экономических явлений и процессов [Текст] : монография. / Н.Я. Бамбаева, А.С. Сорокин – М.: МЭСИ, 2010. – 156 с. – ISBN 978-5-7764-0612-6
2. Ковалев, М., Корженевская, В. Методика построения банковской скоринговой модели для оценки кредитоспособности физических лиц [Текст] // Банки Казахстана. – 1. –2008. – с. 43–48.
3. Ниворожкина, Л.И. Эконометрическое моделирование риска невыплат по потребительским кредитам. [Текст] // Прикладная эконометрика. –30 (2). – 2013. с. 65–76.
4. Сорокин, А.С. К вопросу оценки согласованности мнений экспертов при использовании методов экспертного оценивания в кредитном скоринге. [Текст] /А.С. Сорокин // Роль бизнеса в трансформации общества – 2014: Сб. ст. по мат. IX междунар. научн. конгр. – М.: «Эдитус», 2014. – с. 281-283. – ISBN 978-5-00058-089-9
5. Улитина, Е.В. Статистика: учебное пособие [Текст] / Е.В. Улитина, О.В. Леднева, О.Л. Жирнова – М.: Московский финансово-промышленный университет «Синергия», 2013. – 320 с. – ISBN: 978-5-4257-0107-7
6. Улитина, Е.В. Статистика: учебное пособие [Текст] / Е. В. Улитина, О. В. Леднева, О. Л. Жирнова; под ред. Е. В. Улитиной. - 3-е изд., стер. - Сер. Университетская серия. – М: МФПА, 2011. – 320 с. – ISBN: 978-5-902597-30-8
7. Улитина, Е.В. Применение метода анализа иерархий при согласовании результатов оценки [Текст] / С.В. Харитонов, Е.В. Улитина, В.В. Дик // Прикладная информатика. – 6 (42). – 2012. – с. 108-113
8. Allison, P.D. Logistic regression using the SAS system: theory and application. [Text] – Cary, NC: SAS Institute, 1999. – 303 p. – ISBN 1580253520
9. Anderson R. The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. [Text] – New York: Oxford University press, 2007. – 790 p. – ISBN 0199226407
10. Harrell, Frank. Regression modeling strategies. [Text] – NY: Springer, 2001 – 608 p. – ISBN 0387952322, 9780387952321
11. Hosmer D., Lemeshow S. (1989, 2000, 2013). Applied logistic regression. [Text] – New York: John Wiley and Sons. – 528 p. – 3rd ed. – ISBN 0470582472, 9780470582473
12. Jaccard, J. Interaction effects in logistic regression. [Text] – Thousand Oaks: Sage Publications, 2001. – 70 p. – ISBN 0761922075
13. Kleinbaum, D. G. Logistic regression: A Self-Learning Text. [Text] – New York: Springer-Verlag, 1994. – 282 p. – ISBN 0387941428
14. Lewis, E. M. An introduction to credit scoring. [Text] – San Rafael: The Athena Press, 1992. – 172 p. , – ISBN 9995642239, 978-9995642235
15. Little, R. J. A. A test of missing completely at random for multivariate data with missing values. [Text] // Journal of the American Statistical Association. – 1998. – № 83. – 1198–1202.

16. Lyn C. Thomas. Consumer credit models: pricing, profit, and portfolios. [Text] –New York: Oxford University press, 2009. – 400 p. – ISBN 9780199232130

17. Mays E. (ed.) Handbook of credit scoring. [Text] – Chicago: Glenlake Publishing Company Ltd/Fitzroy Dearborn Publishers, 2001. – 382 p. – ISBN 1888988010, 978-1888988017

18. Naeem, S. Credit risk scorecards: developing and implementing intelligent credit scoring. [Text] – New Jersey: John Wiley and Sons, 2006. – 208 p.– ISBN: 9780471754510

19. Rubin, R. B. Multiple Imputation for Nonresponse in Surveys. [Text] – New York: John Wiley and Sons, Inc., 1987. – 320 p. ISBN 0471655740, 978-0471655749

20. Schafer, J. L. Analysis of Incomplete Multivariate Data. [Text] – London: Chapman and Hall, 1997. – 444 p. – ISBN: 0412040611, 9780412040610

Рецензент: Самойлов Вячеслав Александрович, доктор экономических наук, доцент, и.о. заведующего кафедрой экономики и финансов предприятия Московского финансово-промышленного университета «Синергия».

Alexander Sorokin
Moscow state university of economics, statistics and informatics (MESI)
Moscow university for industry and finance «Synergy»
Russia, Moscow
E-Mail: alsorokin@mail.ru

Building a scorecard using a logistic regression model

Abstract: In banking, credit risk management at one of the key tasks - assessing the creditworthiness of borrowers. The results of the assessment of individual risk are the basis for risk analysis of total loans. Assessment of risk of loan default by the borrower on a particular practice is carried out in two main approaches - based on the subjective opinions of experts or through automated scoring systems.

The basis for constructing the scoring system may take various statistical models. These models may be obtained by linear regression, logistic regression, discriminant analysis , decision trees , neural networks , etc. However, logistic regression is the most commonly used in practice to construct a mathematical model scorecard. The present work is devoted to the different approaches and techniques to build scorecards based on logistic regression, as well as problems that may arise in the construction of scoring models.

The article considers the econometric modeling technique of default probability on the credits on the basis of logistic regression model. The attention is focused on methodical aspects of model creation. The main problems of model creation are illustrated by practical calculations. Transfer technique of received coefficients of logistic regression model to the scorecard is shown. The example of the scorecard creation is given.

Copyright conclusions and recommendations can be used by experts on risk management in commercial banks in constructing scoring systems and checking their work.

Keywords: Credit risk; credit scoring; logistic regression; commercial bank; risk management; scorecard; binning; weight of evidence; information value; validation

Identification number of article 180EVN214

REFERENCES

1. Sorokin, A.S. Primenenie zakonov raspredelenija sluchajnyh velichin dlja modelirovanija jekonomicheskih javlenij i processov [Tekst] : monografija. / N.Ja. Bambaeva, A.C. Sorokin – M.: MJeSI, 2010. – 156 c. – ISBN 978-5-7764-0612-6
2. Kovalev, M., Korzhenevskaja, V. Metodika postroenija bankovskoj skoringovoj modeli dlja ocenki kreditosposobnosti fizicheskih lic [Tekst] // Banki Kazahstana. – 1. –2008. – s. 43–48.
3. Nivorozhkina, L.I. Jekonometricheskoe modelirovanie riska nevyplat po potrebitel'skim kreditam. [Tekst] // Prikladnaja jekonometrika. –30 (2). – 2013. s. 65–76.
4. Sorokin, A.S. K voprosu ocenki soglasovannosti mnenij jekspertov pri ispol'zovanii metodov jekspertnogo ocenivanija v kreditnom skoringe. [Tekst] /A.C. Sorokin // Rol' biznesa v transformacii obshhestva – 2014: Sb. st. po mat. IX mezhdunar. nauchn. kongr. – M.: «Jeditus», 2014. – s. 281-283. – ISBN 978-5-00058-089-9
5. Ulitina, E.V. Statistika: uchebnoe posobie [Tekst] / E.V. Ulitina, O.V. Ledneva, O.L. Zhirnova – M.: Moskovskij finansovo-promyshlennyj universitet «Sinergija», 2013. – 320 c. – ISBN: 978-5-4257-0107-7
6. Ulitina, E.B. Statistika: uchebnoe posobie [Tekst] / E. V. Ulitina, O. V. Ledneva, O. L. Zhirnova; pod red. E. V. Ulitinoj. - 3-e izd., ster. - Ser. Universitetskaja serija. – M: MFPA, 2011. – 320 s. – ISBN: 978-5-902597-30-8
7. Ulitina, E.V. Primenenie metoda analiza ierarhij pri soglasovanii rezul'tatov ocenki [Tekst] / S.V. Haritonov, E.V. Ulitina, V.V. Dik // Prikladnaja informatika. – 6 (42). – 2012. – c. 108-113
8. Allison, P.D. Logistic regression using the SAS system: theory and application. [Text] – Cary, NC: SAS Institute, 1999. – 303 p. – ISBN 1580253520
9. Anderson, R. The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. [Text] – New York: Oxford University press, 2007. – 790 p. – ISBN 0199226407
10. Harrell, Frank. (2001). Regression modeling strategies. [Text] – NY: Springer. – 608 p. – ISBN 0387952322, 9780387952321
11. Hosmer D., Lemeshow S. (1989, 2000, 2013). Applied logistic regression. [Text] – New York: John Wiley and Sons. – 528 p. – 3rd ed. – ISBN 0470582472, 9780470582473
12. Jaccard, J. Interaction effects in logistic regression. [Text] – Thousand Oaks: Sage Publications, 2001. – 70 p. – ISBN 0761922075
13. Kleinbaum, D. G. Logistic regression: A Self-Learning Text. [Text] – New York: Springer-Verlag, 1994. – 282 p. – ISBN 0387941428
14. Lewis, E. M. An introduction to credit scoring. [Text] – San Rafael: The Athena Press, 1992. – 172 p. , – ISBN 9995642239, 978-9995642235
15. Little, R. J. A. A test of missing completely at random for multivariate data with missing values. [Text] // Journal of the American Statistical Association. – 1998. – № 83. – 1198–1202.

16. Lyn C. Thomas. Consumer credit models: pricing, profit, and portfolios. [Text] –New York: Oxford University press, 2009. – 400 p. – ISBN 9780199232130

17. Mays E. (ed.) Handbook of credit scoring. [Text] – Chicago: Glenlake Publishing Company Ltd/Fitzroy Dearborn Publishers, 2001. – 382 p. – ISBN 1888988010, 978-1888988017

18. Naeem, S. Credit risk scorecards: developing and implementing intelligent credit scoring. [Text] – New Jersey: John Wiley and Sons, 2006. – 208 p.– ISBN: 9780471754510

19. Rubin, R. B. Multiple Imputation for Nonresponse in Surveys. [Text] – New York: John Wiley and Sons, Inc., 1987. – 320 p. ISBN 0471655740, 978-0471655749

20. Schafer, J. L. Analysis of Incomplete Multivariate Data. [Text] – London: Chapman and Hall, 1997. – 444 p. – ISBN: 0412040611, 9780412040610