УДК 519.866

ОЦЕНИВАНИЕ ВЕРОЯТНОСТИ ДЕФОЛТА ПО КРЕДИТНЫМ ОПЕРАЦИЯМ С ИСПОЛЬЗОВАНИЕМ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ И КЛАСТЕРНОГО АНАЛИЗА

С.С. СЕРЕДНИЙ

Предложен метод оценки вероятности дефолта по кредитным операциям с применением логистической регрессии и кластерного анализа. Приведен пример применения данного метода на реальной выборке, на которой предложенный метод показал большую эффективность по сравнению с методом, основанном только на логистической регрессии.

ВВЕДЕНИЕ

Основную часть своего дохода банки получают за счет кредитной деятельности, что делает постоянную разработку и совершенствование методов оценки вероятности дефолта по кредитным операциям актуальной для банковского сектора. Следует отметить, что результат кредитной операции имеет случайную природу и зависит от непрогнозируемых и сложнопрогнозируемых форс-мажорных факторов (смерть заемщика, неурожай для сельхоз предприятий), поэтому на момент выдачи кредита отнести заемщика однозначно к «плохим» или «хорошим» не представляется возможным.

Существует два основных подхода к оценке вероятности дефолта заемщика. Согласно первого подхода дефолт заемщика представляется как превышение суммы задолженности заемщика над рыночной стоимостью его активов, и на основании истории изменения биржевых цен на акции заемщика рассчитывается вероятность падения их ниже суммы задолженности. Основным представителем данного подхода является модель KMV [1]. Но такой подход имеет ряд существенных недостатков: он неприменим для оценки вероятности дефолта физических лиц и небольших компаний, а также его нельзя применить в украинских условиях, так как в Украине фактически отсутствует ликвидный фондовый рынок.

По второму подходу задача оценки вероятности дефолта интерпретируется как задача классификации заемщиков на «плохих» и «хороших» (или, по необходимости, на большее количество классов). В рамках данного подхода, в свою очередь, существует два варианта решения данной задачи: на основании четкой и нечеткой классификации заемщиков. В литературе

встречается описание использования следующих методов для оценки вероятности дефолта по кредитным операциям:

- На основании четкой классификации — кластерный анализ [2], деревья решений [3], нейронные сети [4].
- На основании нечеткой классификации — наивный байесовский подход [5], логистическая регрессия [6].

Согласно первому варианту обучающую выборку разбивают на классы по возможности таким образом, чтобы в каждом классе находились представители только одной категории. Вероятность дефолта оценивается как доля «плохих» заемщиков в каждом из классов, на которые разбита обучающая выборка. Такой подход к решению данной задачи является недостаточно корректным, поскольку не учитывает случайную природу результата кредитной операции.

Согласно второму варианту оценивается вероятность принадлежности заемщика к одной из категорий, при этом категорий обычно берут две («плохую» или «хорошую»), однако исходя из потребностей банка их количество может расширяться (например, «имел просрочки свыше 90 дней»). Но наивный байесовский подход имеет существенный недостаток, который заключается в том, что он построен на «наивной» гипотезе о том, что параметры, описывающие заемщика, независимы между собой, хотя это в действительности не так: например, возраст заемщика существенно коррелирует с такими параметрами, как имущественное состояние, социальный статус и т.д.

Цель работы — по имеющимся данным о кредитных операциях, содержащих операцию о параметрах кредита и сведения о заемщике, построить эффективную модель оценки вероятности дефолта по кредитным операциям новых заемщиков.

МОДИФИЦИРОВАННАЯ МОДЕЛЬ ОЦЕНКИ ВЕРОЯТНОСТИ ДЕФОЛТА

Оценка вероятности дефолта методом логистической регрессии происходит исходя из двух основных предположений:

- Исход кредитной операции зависит от ненаблюдаемой величины $\hat{Y} = X\beta + \beta_0 + \varepsilon.$ При этом, если ненаблюдаемая величина $\hat{Y} > 0$, то считается, что клиент не погасит кредит, а если $\hat{Y} \le 0$, то погасит.
- Независимая случайная величина имеет логистическое распределение, имеющее функцию распределения $F(x) = \dfrac{1}{1+e^{-x}}$.

Исходя из этих предположений, вероятность дефолта рассчитывается как

$$PD = P(\varepsilon > -X\beta - \beta_0) = 1 - P(\varepsilon \le -X\beta - \beta_0) = 1 - \frac{1}{1+e^{X\beta+\beta_0}} = \frac{1}{1+e^{-X\beta-\beta_0}}.$$

Исходя из особенностей построения логистической регрессии, можно выделить два основных недостатка, ухудшающих точность оценки вероятности дефолта.

- Логистическая регрессия не предназначена для обработки качественных параметров, и их приходится заменять на числовые. При этом теряется их информативность.

- Логистическая регрессия не учитывает взаимосвязи между переменными и наличие «границ чувствительности» для некоторых параметров (например, разница в один год для заемщиков в возрасте 60 и 61 лет намного более существенна, чем для заемщиков в возрасте 30 и 31 года).

Для минимизации негативного влияния первого недостатка в работе [7] было предложено использовать показатель WOE (Weight Of Evidence — вес доказательства) при замене качественных параметров на числовые. Показатель WOE, который рассчитывается по формуле: $WOE_i = \ln\left(\frac{G_i}{B_i}\right)$, где G_i — доля «хороших» заемщиков от общего числа «хороших» заемщиков, для которых категориальный параметр принимает *i*-тое значение; B_i — доля «плохих» заемщиков от общего числа «плохих», для которых категориальный параметр принимает *i*-тое значение.

Второй недостаток логистической регрессии следует из того, что ненаблюдаемая величина должна иметь нелинейный вид, поскольку зависимость результата кредитной операции от параметров не является линейной (отсутствие собственного жилья в 55 лет более рискованно, чем в 25 лет, а возраст 30 лет менее рискованный, чем возраст 20 или 60 лет), поэтому представление ненаблюдаемой величины в виде $\hat{Y} = X\beta + \beta_0 + \varepsilon$ вносит ошибку, связанную с игнорированием слагаемых более высоких порядков.

При простом включении слагаемых более высоких порядков, мы рискуем потерять точность модели за счет эффекта «подстраивания», поскольку добавление одной переменной приводит к существенному увеличению требований к количеству входящих данных. Эта проблема особенно актуальна при решении задачи оценки вероятности дефолта, поскольку данная задача часто решается в условиях маленькой обучающей выборки.

Исходя из приведенных соображений, для минимизации негативного влияния второго недостатка логистической регрессии был предложен следующий подход:

- обучающая выборка при помощи кластерного анализа разбивается на кластеры в соответствии с подобностью параметров, которые описывают кредитную операцию;
- в каждом из полученных кластеров независимо от других строится модель оценки вероятности дефолта по кредитной операции на основании логистической регрессии;
- для нового заемщика сначала определяется кластер, в который он входит, а далее оценивается вероятность дефолта при помощи модели, построенной для данного кластера.

МОДЕЛИРОВАНИЕ РЕЗУЛЬТАТОВ КРЕДИТНЫХ ОПЕРАЦИЙ НА РЕАЛЬНЫХ ДАННЫХ

Для практической проверки эффективности предлагаемого метода были построены модели на основании стандартного метода логистической регрессии, а также на основании логистической регрессии и кластерного анализа.

Моделирование проводилось на основании выборки, предоставленной компанией SAS Institute (международная компания, являющаяся одним из лидеров рынка разработки программного обеспечения, в частности, в области риск-менеджмента). Выборка содержит данные про 2102 кредитные операции с указанием их результата («дефолт» и «не дефолт») и 29 параметров, характеризирующих заемщика и кредитную операцию.

Алгоритм построения модели оценки вероятности дефолта заемщика по кредитной операции состоит из следующих этапов:

- предварительный выбор параметров на основании представлений о предметной области (например, ФИО заемщика и его ИНН не влияют на результат кредитной операции) и замена абсолютных параметров на относительные (сумма кредита и среднемесячный доход заемщика сами по себе не информативны и их следует заменить на соотношение среднемесячного платежа по кредиту к среднемесячным доходам);
- удаление из выборки кредитов, полученных мошенниками (поскольку такие кредитные операции имеют другую зависимость между параметрами кредитной операции и ее результатами);
- очистка выборки от аномальных данных, ошибок и логических ошибок;
- обработка пропущенных значений;
- группирование значений качественных параметров, которые редко встречаются (типы товаров, профессии заемщиков);
- разбиение выборки на обучающую и тестовую;
- разбиение выборки на кластеры, дальнейшие этапы проводятся для каждого из полученных кластеров независимо;
- группирование значений качественных параметров, которые редко встречаются по итогам разбиения на кластеры;
- замена всех качественных параметров на числовые, для чего используется показатель WOE [7];
- нормализация всех параметров для обеспечения устойчивости работы программной реализации алгоритма;
- расчет статистической значимости параметров и корреляции между ними, а также группирование параметров, имеющих высокую корреляцию с целью избежать при построении модели ложных корреляций.

С целью определения оптимального количества параметров, которые стоит включать в модель, производятся следующие этапы:

- сортировка параметров по их статистической значимости;
- повторное разбиение выборки на обучающую и тестовую;
- поиск коэффициентов логистической регрессии при включении в модель только одного параметра, имеющего наибольшую статистическую значимость;
- поиск коэффициентов логистической регрессии при включении в модель следующего по статистической значимости параметра. Данный этап повторяется до тех пор, пока новая модель (т.е. с количеством параметров увеличенным на один) будет иметь большую эффективность на тестовой выборке, чем старая.

Добавление в модель параметров по одному не только позволит определить то оптимальное количество параметров, после которого модель начинает терять точность и начинает попросту «подстраиваться» под обучающую выборку, но еще и существенно ускоряет и повышает устойчивость работы программного алгоритма. В случае, если на каждой новой итерации в качестве начальных значений выбирать оптимальные значения, рассчитанные на предыдущем шаге и 0 для вновь добавленного параметра, то алгоритм поиска значений для логистической регрессии, например, с десятью параметрами, будет работать быстрее, чем просто поиск сразу всех десяти значений из произвольных точек.

В рамках построения численного эксперимента обучающая выборка разбивалась на 2 и 3 кластера. Эффективность модели с разбиением обучающей выборки на 3 кластера оказалась хуже, чем с разбиением на 2 кластера, что связано, по всей видимости, с недостаточным объемом выборки.

ОПИСАНИЕ И СРАВНЕНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

Для анализа полученных результатов была использована ROC-кривая [6], смысл которой заключается в том, что она показывает зависимость соотношения количества верно классифицированных «хороших» заемщиков к неверно классифицированным «плохим» заемщикам в зависимости от порога отсечения. В качестве порога отсечения выступает рассчитанная вероятность дефолта *PD*. Построение ROC-кривой происходит следующим образом: по оси *0Y* откладывается процент правильно классифицированных «хороших» заемщиков, а по оси *0X* количество неправильно классифицированных «плохих» заемщиков в зависимости от значения порога отсечения. Соответственно, чем больше ROC-кривая отклоняется от диагональной линии, тем более эффективной считается построенная модель. Сама диагональная линия считается абсолютно бессмысленным классификатором, который эквивалентен произвольному выбору.

Для упрощенного сравнения эффективности моделей используется показатель *Gini*, представляющий из себя удвоенную площадь между диагональной линией и ROC-кривой.

Коэффициент *Gini* рассчитывается по формуле

$$Gini = 2\left(\int_0^1 ROC(x)dx - 0,5\right).$$

В результате проведенного моделирования были получены следующие результаты: для модели на основании стандартного метода логистической регрессии — показатель $Gini = 0,629042$, а для модели на основании логистической регрессии с применением кластерного анализа — показатель $Gini = 0,654799$. Графики ROC-кривых, полученные для обоих подходов, приведены ниже.

Как видно на рисунке, а также из значений показателя *Gini* для обеих моделей, модель с применением кластерного анализа более эффективна, чем модель на основании стандартного метода логистической регрессии.

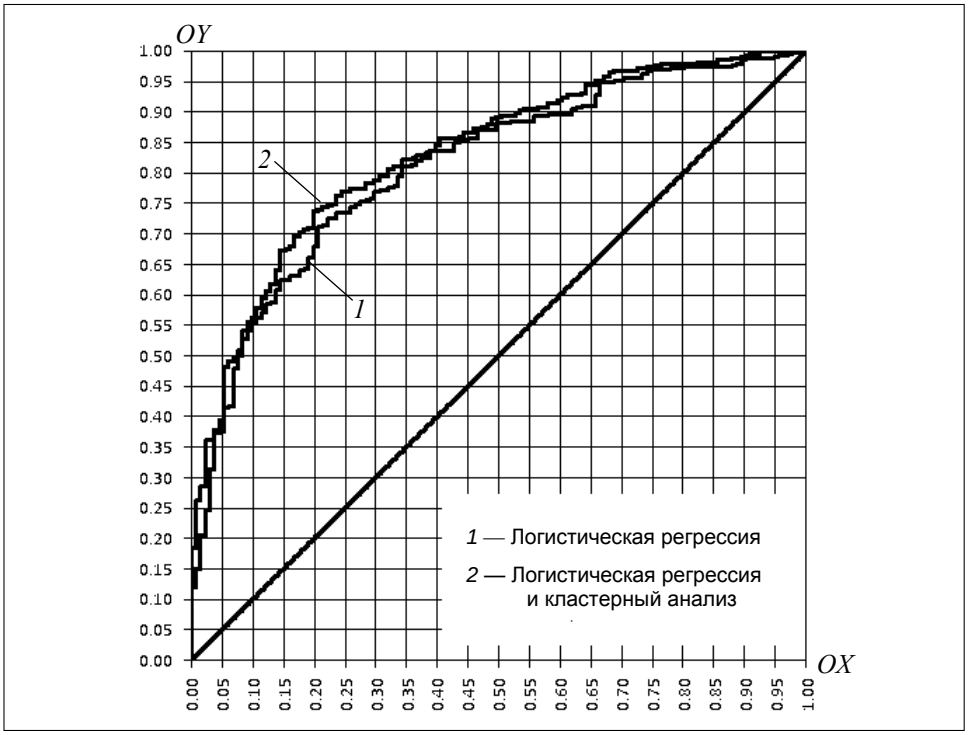

Рисунок. Пример графиков ROC-кривых для рассматриваемых методов оценки вероятности дефолта

ВЫВОДЫ

За счет разбиения входящей выборки на кластеры по принципу однородности удалось достичь уменьшения влияния на точность оценки того недостатка логистической регрессии, что она не учитывает взаимосвязи между параметрами и наличия границ чувствительности. Приведенный в работе подход к оценки вероятности дефолта по кредитной операции на основе логистической регрессии и кластерного анализа показал большую эффективность, чем подход на основании стандартного метода логистической регрессии, о чем свидетельствуют результаты полученные на тестовом примере.

Использование разработанного подхода оценки вероятности дефолта по кредитной операции позволит улучшить кредитные портфели банков, что, в свою очередь, позволит увеличить прибыльность и устойчивость банковского сектора в целом. К тому же, при увеличении объема входящей выборки можно ожидать дальнейшего увеличения точности модели, так как это позволит разбить обучающую выборку на большее количество кластеров.

В предлагаемом подходе в дальнейшем может быть реализована разработка четких критериев, а также алгоритма определения оптимального количества кластеров и параметров, по которым должно происходить разбиение входящей выборки.

ЛИТЕРАТУРА

1. *Peter J. Crosbie, Jeffrey R. Bohn.* Modeling Default Risk. — 2002. — http:// www.creditrisk.ru/publications/ files_attached/modeling_default_risk.pdf.
2. *Оценка* вероятности банкротства предприятий-заемщиков на основе кластерного анализа // Экономический анализ: теория и практика. — 2007. — № 18. — С. 44–45.
3. *Ларин С., Ходжаева И.* Использование деревьев решений для оценки кредитоспособности физических лиц // Банковское дело. — 2004. — № 3. — С. 30–33.
4. *Лаврушин О.И., Афанасьева О.Н., Корниенко С.Л.* Банковское дело: современная система кредитования. — М.: Кнорус, 2007. — 261 с.
5. *Воронцов К.В.* Лекции по статистическим (байесовским) алгоритмам классификации. — 2008. — www.ccas.ru/voron/download/Bayes.pdf.
6. *Палкин Н.* Логистическая регрессия и ROC-анализ — математический аппарат. — http://www.basegroup.ru/library/analysis/regression/logistic/.
7. *Ковалев М., Корженевская В.* Методика построения банковской скоринговой модели для оценки кредитоспособности физических лиц. — www.bsu.by/ ru/sm.aspx?guid=49623.

Поступила 11.02.2010