Тлеппаев А.М., Бушмелев И.В.

Применение модели бинарного выбора для скоринга в банковской деятельности

Данная статья посвящена описанию модели бинарного выбора и обоснованию ее выбора для построения скоринговой модели в банках. Из множества моделей, которые способны определить влияние факторов на одну переменную, здесь рассматривалась именно логистическая регрессия. Именно она является традиционным статистическим инструментом для расчета коэффициентов скоринговой карты, а ROC-анализ обеспечивает управление рисками в зависимости от кредитной политики и стратегии организации. Логистическая регрессия предназначена для обширного спектра функций, в том числе анализа связи между определенным количеством независимых переменных и зависимой переменной. Данная логистическая регрессия является бинарной, что обозначает то, что зависимая переменная может принимать только два значения. Иными словами логистическая регрессия помогает оценить вероятность того, что некое событие наступит или не наступит для конкретного случая, в нашем варианте это возврат кредита или же дефолт. По данным результатам можно построить зависимость между поведением клиента и его платежеспособностью, и в последующем применять данную модель в банках при выдаче займа.

Ключевые слова: скоринг, дефолт, логистическая регрессия, модель бинарного выбора, банк.

Tleppaev A.M., Bushmelev I.V.

Application of a binary choice model for bank scoring

This article describes a binary choice model and justification of its choice for the construction of the scoring model in banks. Among the many models that are able to determine the influence of factors on a single variable, in this very case the logistic regression was considered. That is exactly the logistic regression that is the traditional statistical tools to calculate coefficients scorecard, and ROC-analysis provides a risk management, depending on the credit policy and strategy of the organization. Logistic regression is used for a wide range of functions, including the analysis of the dependence between certain number of independent variables the dependent variable. This is a binary logistic regression, which means that the dependent variable can take only two values. In other words logistic regression helps to assess the probability that an event occurs or does not occur for a particular case, in our case, a return of the loan or default. According to the results, you can build a dependence between the behavior of the client and his ability to pay, and subsequently apply this model to the banks when issuing loans.

Key words: scoring, default, logistic regression, binary choice model, bank.

Тлеппаев А.М., Бушмелев И.В.

Банк қызметінің скорингі үшін бинарлық моделді қолдану

Бұл мақала банктерде скоринг моделін ендіру үшін бинарлық іріктеу үлгісін сипаттауға және оны таңдау негіздемесіне арналған. Бұл жерде көптеген модельдердің ішінен бір ауыспалыға факторлардың әсер етуін анықтай алатын логистикалық регрессия қарастырылды. Ол скорингтік сызба коэффицентін есептейтін дәстүрлік статистикалық құралы болып саналады, ал ROC-талдау ұйымның кредиттік саясаты мен стратегиясына байланысты тәуекелдерді басқарумен қамтамасыз етеді. Логистикалық регрессия қызметтердің кең ауқымды түрлеріне арналған, сонын ішінде тәуелді және тәуелсіз ауыспалылардың белгілі бір көлемі арасындағы қатынастарды талдауы да бар. Осы логистикалық регрессия бинарлы болып табылады, бұл тәуелді ауыспалының тек қана екі мәнді қабылдай алатының білдіреді. Басқаша айтқанда логистикалық регрессия нақты оқиға үшін белгілі бір жағдайдың болу немесе болмау мүмкіндігін бағалауға көмектеседі. Нәтижесінде клиенттің мінез-құлқы мен төлеу қабілеті арасындағы байланысын құруға болады және сонымен қатар осы модельді банктерде заем берер кезінде қолдануға болады.

Түйін сөздер: скоринг, логистикалық регресс, бинарлық іріктеу моделі, банк.

УДК 336.71

*Тлеппаев А.М., Бушмелев И.В.

Казахстанско-Немецкий университет,
Республика Казахстан, г. Алматы
*E-mail: arsentlp@gmail.com

ПРИМЕНЕНИЕ МОДЕЛИ БИНАРНОГО ВЫБОРА ДЛЯ СКОРИНГА В БАНКОВСКОЙ ДЕЯТЕЛЬНОСТИ

Логистическая регрессия является одним из важных инструментов для решения задач регрессии и классификации. Данный метод является неотъемлемым атрибутом в медицине и с относительно недавнего времени начал применяться в скоринге для оценки заемщиков. Логистическая регрессия предназначена для обширного спектра функций, в том числе анализ связи между определенным количеством независимых переменных и зависимой переменной. Данная логистическая регрессия является бинарной, что обозначает то, что зависимая переменная может принимать только два значения. Иными словами логистическая регрессия помогает оценить вероятность того, что некое событие наступит или не наступит для конкретного случая, в нашем варианте это возврат кредита или же дефолт.

Пусть рассматривается исход по займу, тогда задается переменная У с двумя значениями 1 или 0, где 1 – клиент расплатился по кредиту, 0 – дефолт.

Но здесь возникает некая неопределенность из-за бинарной природы переменной и модель будет иметь предсказанные значения превышающие 1 или меньше нуля. Однако значения превышающие данные пределы недопустимы для первоначальной задачи. Чтобы разрешить сложившуюся ситуацию, необходимо иначе сформулировать задачу регрессии. Вместо предыдущего предсказания зададим непрерывную переменную со значениями на отрезке от 0 до 1 при любых значениях независимых переменных [1]. Данное преобразование осуществляется при помощи уравнения логит-преобразования:

$$P = \frac{1}{1 + e^{-y}} \quad (1)$$

где Р – вероятность того, что произойдет интересующее событие; е-экспонента – основание натурального логарифма ≈2,71; у – стандартное уравнение регрессии.

График зависимости между вероятностью события и величины объясняемой переменной показан на рисунке 1.

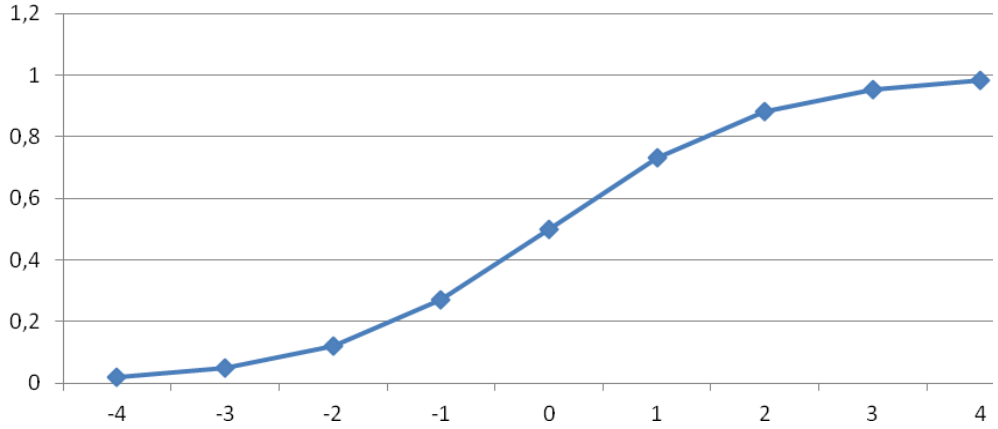

Рисунок 1 – Логистическая кривая (составлено автором с использованием источника [1])

Преобразуем вероятность P следующим образом:

$$P^{\zeta} = \log_{e} \frac{P}{(1-P)} \qquad (2)$$

Данное преобразование имеет название – логистическое или же логит-преобразование. Теоретически P^{ζ} может принимать любое значение.

Есть несколько способов, с помощью которых представляется возможным найти коэффициенты логистической регрессии. Один из них метод максимального правдоподобия.

С помощью данного метода возможно получить оценки параметров генеральной совокупности по некоторым данным выборки. Функция правдоподобия, или как ее еще называют likehood function – L, выражает плотность вероятности совместного появления результатов выборки.

Исходя из логики данного метода, в качестве оценки независимого параметра принимается такое значение, которое обеспечивает реализацию следующего условия $L \to \infty$.

Оценку независимого параметра можно произвести значительно удобнее, если максимизировать не саму функцию *L,* а натуральный логарифм от нее. Это представляется логичным, поскольку максимальное значение обеих функций достигается при одинаковом значении θ:

$$L*(Y,\theta) = \ln(L(Y,\theta)) \to max. \qquad (3)$$

В том случае, если используется бинарная независимая переменная, то *Pi* – вероятность появления единицы *Prob(Yi = 1)* зависящая от *XiW*, где *Xi* – строка матрицы регрессоров, *W* – вектор коэффициентов регрессии:

$$P_i = F(X_i W), F(z) = \frac{1}{1+e^{-z}} \qquad (4)$$

Из вышесказанного очевидно, что логарифмическая функция правдоподобия тождественна:

$$L^* = \sum_{i \in I1} \ln P_i(W) + \sum_{i \in I0} \ln(1 - P_i(W)) = \sum_{i=1}^{k} \left(Y_i \ln P_i(W) + (1 - Y_i) \ln[(1 - P_i(W))]\right) \qquad (5)$$

где I0, I1 – множества наблюдений, для которых *Yi=0* и *Yi = 1* соответственно.

Логистическая регрессия не может моделировать нелинейные зависимости , однако чтобы оценить качество данной модели, можно применить ROC – анализ – эффективный инструмент для оценки качества моделей логистической регрессии, а чтобы рассчитать коэффициенты логистической регрессии можно применить любой из градиентных методов, такие как: метод сопряженных градиентов, методы переменной метрики и другие.

ROC-кривая или Receiver Operator Characteristic – название которое произошло из

систем обработки сигналов – кривая, которую часто используют, чтобы представить результаты бинарной классификации.

Из вышесказанного очевидно, что классов для объясняемой переменной два, один из них называется классом с положительным исходом, другой соответственно с отрицательным. ROC – кривая дает понять зависимость количества положительных примеров, которые были верно классифицированы от количества отрицательных примеров, которые были неверно классифицированы. Те положительные примеры, которые были классифицированы верно, называются истинно положительными, исходя из терминов ROC-анализа, а те отрицательные примеры, которые были классифицированы неверно – ложно отрицательные. Пусть у классификатора имеется некоторый параметр, называемый точкой отделения, с помощью которого осуществимо то или иное разбиение на два класса, всего лишь варьируя данный показатель. В зависимости от этого параметра будут получаться различные величины ошибок первого и второго рода [2].

В логистической регрессии порог отсечения находится в пределах 0 – 1 – еще это называют расчетным значением уравнения регрессии или рейтингом.

В таблице 1 приведена расшифровка результатов классификации модели:

- TP (True Positives) – положительные примеры, верно классифицированные или истинно положительные случаи;
- TN (True Negatives) – отрицательные примеры, верно классифицированные или истинно отрицательные случаи;
- FN (False Negatives) – положительные примеры, ложно классифицированные, т.е. те положительные, которые классифицированы как отрицательные – ошибка первого рода. Еще этот случай называют «ложный пропуск» – когда интересующее событие не обнаруживается по ошибке (ложно отрицательные примеры)
- FP (False Positives) – отрицательные примеры, ложно классифицированные, т.е. отрицательные примеры, которые классифицировались как положительные – это называется «ложное обнаружение», т.е. когда события нет, но решается, что оно имеет место быть из-за допуска ошибки, или ложно положительные случаи.

Таблица 1 – Взаимосвязь критериев классификации модели [1]

	Фактически	
Модель	Положительно	Отрицательно
Положительно	TP	FP
Отрицательно	FN	TN

Присвоение событию отрицательный или положительный характер зависит от той задачи, которую предстоит решить. Например, если прогнозируется, что есть вероятность того, что клиент не отдаст кредит, то положительным исходом события будет класс «дефолт клиента», отрицательным «надежный клиент». И наоборот, если прогнозируется что клиент добросовестный, то положительным исходом будет возврат клиентом кредита, а отрицательным – дефолт клиента.

При анализе данных чаще всего используются относительные, но не абсолютные показатели – доли, выраженные в процентах. Так, % доля истинно положительных примеров, или True Positives Rate, выглядит следующим образом:

$$TPR = \frac{TP}{TP + FN} * 100\% \quad (6)$$

Доля ложно положительных примеров или False Posotives Rate:

$$FPR = \frac{FP}{TN + FP} * 100\% \quad (7)$$

Специфичность и чувствительность модели так же являются важными определениями. С помощью данных показателей можно определить объективную ценность любого бинарного классификатора.

Чувствительность или Sensitivity – доля верно классифицированных – истинно положительных примеров:

$$S_e = TRP = \frac{TP}{TP + FN} * 100\% \quad (8)$$

Специфичность (Specificity) – доля отрицательных примеров верно классифицированных или доля истинно отрицательных случаев, которые были правильно идентифицированы моделью:

$$S_p = \frac{TN}{TN + FP} * 100\% \quad (9)$$

Необходимо заметить, что FPR = 100-S_p.

Модель, которая обладает высокой чувствительностью, часто дает истинной результат, если имеется положительный исход (обнаружены положительный случаи). Модель же с высокой специфичностью чаще дает истинный результат, если имеется отрицательный исход (обнаружены отрицательные примеры).

Для любого значения порога отсечения, которое изменяется от 0 до 1 с шагом dx (например, 0,01) в ROC-кривой рассчитываются значения чувствительности Se и специфичности Sp. Каждое последующее значение примера в выборке может являться порогом в качестве альтернативы. Далее строится график зависимости, где чувствительность Se откладывается по оси Y, а по оси X откладывается 100% – Sp (сто процентов минус специфичность) или тождественное оному выражение FRP – доля ложно положительных случаев [3]. Данный график часто дополняется прямой $y = x$.

Необходимо так же иметь в виду, что допустим и имеется способ расчета точек ROC-кривой, который является более экономичным, чем пример, приведенный выше. Его экономичность обуславливается тем, что его сложность вычисления является нелинейной и для каждого порога следует каждый раз в каждой записи рассчитывать TP и FP. Двигаясь по набору данных в обратном направлении, которое отсортировано по убыванию выходного поля классификатора (рейтингу), таким образом можно за один проход вычислить значения всех точек ROC-кривой, последовательно подвергая обновлению значения TP и FP.

Для идеального классификатора график ROC-кривой проходит через верхний левый угол, где доля истинно положительных случаев составляет 100%, или 1,0, что характеризует идеальную чувствительность, а доля ложно положительных примеров, напротив, равна нулю. Посему чем кривая находится ближе к верхнему левому углу, тем больше вероятности предсказания моделью. И напротив, чем кривая является менее изогнутой и чем более ближе она расположена к диагональной прямой, тем модель является менее эффективной. По сути диагональная прямая $x = y$ является моделью бесполезности и соответствует абсолютно нефункционирующему классификатору, иными словами модель не может различить два класса друг от друга.

Если провести некую визуальную оценку ROC-кривой, то расположения ее комбинаций относительно между собой говорит о их сравнительной эффективности. Кривая, что расположена выше и левее, говорит о большей предсказательной способности модели. Однако стоит отметить тот факт, что не смотря на кажущуюся простоту, визуальная оценка не всегда позволяет с точностью определить модель, чьи предсказательные способности являются наиболее эффективными. Существует более точный способ оценить данный показатель. Он представляет собой геометрический смысл и заключается в оценках площади под ROC-кривыми в процессе их сравнения [4].

Теоретически модель изменяется от 0 до 1, но так как модель логичнее оценивать и характеризовать кривой, которая расположена выше положительной диагностики, то обычно следует обращать внимание только лишь на изменения от 0,5 до 1 – идеальная модель. Эта оценка может быть получена различными способами в том числе и непосредственно вычисляя площади под многогранником, что ограничен осями координат справа и снизу, и точками, полученными в результате эксперимента слава вверху. Численный показатель площади носит название AUC (Area Under Curve).

Очень грубо можно считать, что показатель AUC прямо пропорционален прогностической силой, присущей модели, но так же следует иметь в виду, что данный показатель нужен, когда сравнивают несколько моделей, а так же AUC не содержит информации о таких важных показателях, как чувствительность и спецификация модели.

В таблице 2 приведена шкала значений AUC, которая призвана оценить качество модели. Идеальная модель будет обладать стопроцентной чувствительностью и специфичностью, но, конечно же, в реальности добиться таких показателей не представляется возможным. Мож-

но сказать даже, что невозможно одновременно повысить такие показатели, как чувствительность и специфичность модели. Порог отсечения дает нам возможность достижения некоего компромисса. Пороговое отсечение влияет на соотношение показателей Se и Sp, и в данном случае необходимо найти некий оптимальный порог отсечения.

Таблица 2 – Шкала значений AUC [1]

Интервал AUC	Качество модели
0,9 – 1,0	Отличное
0,8 – 0,9	Очень хорошее
0,7 – 0,8	Хорошее
0,6 – 0,7	Среднее
0,5 – 0,6	Неудовлетворительное

Данная процедура с определением оптимального порога отсечения необходима для того, чтобы применить модель на практике, т.е. относить новые параметры к одному из двух классов. Очевидно, что для того, чтобы определить оптимальный порог (optimal cut of value), необходимо задать некий критерий, по которому он будет определяться. Это является очень важным, каждая задача диктует свою оптимальную стратегию;

- требование минимальной величины чувствительности или специфичности модели;

- требования максимальной суммарной чувствительности и специфичности модели:

$$\mathbf{Cut_{off_0}} = \max_{\mathbf{k}}(\mathbf{Se_k} + \mathbf{Sp_k}) \qquad (10)$$

- требования баланса между чувствительностью и специфичностью, иными словами тот случай, когда Se приблизительно равно Sp:

$$\mathbf{Cut_{off_0}} = \min_{\mathbf{k}}|\mathbf{Se_k} - \mathbf{Sp_k}| \qquad (11)$$

Приведенные выше требования могут выступать критериями выбора порогового отсечения. Следующее значение порога предлагается по умолчанию, а в третьем случае порог является точкой пересечения двух кривых. По оси X откладывается порог отсечения, а по У – чувствительность или специфичность модели. Таким образом, достигается точка баланса между чувствительностью и специфичностью.

Логистическая регрессия является традиционным статистическим инструментом для расчета коэффициентов скоринговой карты. А ROC-анализ обеспечивает управление рисками в зависимости от кредитной политики и стратегии организации.

В рамках политики банка модели ставится задача выявления неблагонадежных потенциальных заемщиков. Но поскольку в скоринге общепринято, что чем выше рейтинг клиента, тем выше его кредитоспособность, то считается положительным исходом успешное погашение займа, а отрицательным – дефолт по кредиту.

Исходя из этого, можно заключить, что скоринговая модель с высокой специфичностью соответствует консервативной кредитной политике (чаще происходит отказ в выдаче кредита), а с высокой чувствительностью – политике рискованных кредитов. В первом случае минимизируется кредитный риск, связанный с потерями ссуды и процентов и дополнительными расходами на возвращение кредита, а во втором – коммерческий риск, связанный с упущенной выгодой.

Литература

1 Цыплаков А.А. Некоторые эконометрические методы Метод максимального правдоподобия в экономике. – М: 2011. – 100 с.

2 Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. Kluwer Academic Publishers: 2004. – 85 с.

3 Zweig M.H., Campbell G. ROC Plots: A Fundamental Evaluation Tool in Clinical Medicine // Clinical Chemistry. – 1993. – Vol. 39. – No. 4. – P. 22-27.

4 Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves // Proc. of 23 International Conference on Machine Learning. – Pittsburgh, PA, 2006. – P. 14-18.

References

1 Cyplakov A. A. Nekotorye jekonometricheskie metody. Metod maksimalnogo pravdopodobija v ekonometrike. – M: 2011. – 100 s.

2 Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. Kluwer Academic Publishers: 2004. – 85 p.

3 Zweig M.H., Campbell G. ROC Plots: A Fundamental Evaluation Tool in Clinical Medicine // Clinical Chemistry. – 1993. – Vol. 39. – No. 4, – P. 22-27.

4 Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves // Proc. of 23 International Conference on Machine Learning. – Pittsburgh, PA, 2006. – P. 14-18.