

Отраслевым стандартом оценки эффективности бинарных классификаторов в банковском моделировании можно считать использование коэффициента Gini и AUC-ROC. При этом практически все выборки (портфели), на которых строятся банковские модели, являются несбалансированными, то есть объектов одного класса (например, дефолтных кредитов) значительно меньше, чем объектов другого класса (недефолтных кредитов). Попробуем разобраться, соответствует ли бизнес-целям банка применение Gini и AUC-ROC в моделировании, и предложим альтернативные метрики для оценки качества банковских моделей.

Сергей АФАНАСЬЕВ, КБ «Ренессанс Кредит» (ООО), исполнительный директор, начальник управления расследования мошенничества

Анастасия СМЕРНОВА, КБ «Ренессанс Кредит» (ООО), главный эксперт по работе с системами противодействия мошенничеству

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

Бинарные классификаторы являются, пожалуй, наиболее популярными методами машинного обучения, применяемыми на практике.

В банковской сфере модели бинарной классификации используются практически во всех ключевых направлениях деятельности:

- в скоринге это модели прогнозирования вероятности дефолта заемщика;
- в перекрестных продажах (CRM) — модели отклика клиента на x-sell предложение;
- во взыскании — модели прогнозирования вероятности возврата просроченного кредита;
- в антифроде — модели выявления мошеннических кредитов;
- и т.д.

Бинарная классификация используется в задачах, где объекты выборки делятся на два класса — положительные (positive) и отрицательные (negative) (рис. 1а). Применительно к банковским задачам это могут быть следующие разбиения:



Сергей АФАНАСЬЕВ Анастасия СМИРНОВА

- попадет кредит в просрочку (1 — positive) или нет (0 — negative);
- откликнется клиент на x-sell предложение банка (1) или нет (0);
- вернет клиент просроченный кредит (1) или не вернет (0);
- окажется кредитная заявка мошеннической (1) или нет (0);
- и т.д.

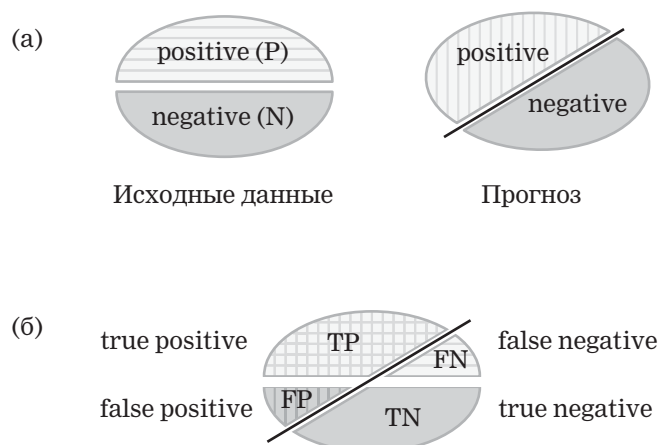
В свою очередь сама модель бинарной классификации присваивает объектам также две метки — positive или negative (см. рис. 1а). А поскольку моделям свойственно ошибаться, то в результате бинарной классификации все объекты выборки разбиваются на четыре типа, образуя матрицу ошибок (confusion matrix)¹ (рис. 1б):

- 1) истинно положительные (true positive — TP);
- 2) истинно отрицательные (true negative — TN);
- 3) ложно положительные (false positive — FP) — ошибка 1-го рода (type I error);
- 4) ложно отрицательные (false negative — FN) — ошибка 2-го рода (type II error).

Метрики качества бинарных классификаторов делятся на «пороговые» (single-threshold) и «не зависящие от порога» (threshold-free).

Рисунок 1

Исходные и прогнозные данные бинарной классификации и матрица ошибок



¹ Дословно confusion matrix может быть переведена как «матрица путаницы», что имеет некоторый скрытый смысл, поскольку часто возникает путаница при определении ошибок 1-го и 2-го рода.

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

Бинарные классификаторы обычно настраиваются на прогноз вероятности принадлежности к одному из двух классов. То есть каждому объекту присваивается числовое значение в диапазоне от 0 до 1 (или от 0 до 999 для скорингового балла). Таким образом, чтобы разделить прогнозные значения на два класса, необходимо выбрать пороговую вероятность, разбивающую весь диапазон значений вероятности на две группы. После выбора пороговой вероятности можно построить матрицу ошибок, на основе которой рассчитываются различные «пороговые» метрики качества.

Если переменные в матрице ошибок рассчитать для различных пороговых вероятностей, то можно построить метрики качества, не зависящие от порога, такие как Gini, AUC-ROC и др. Разберем несколько полярных метрик качества, применяемых в машинном обучении.

Пороговые метрики качества

1. *Accuracy* (правильность¹):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

Метрика *accuracy* показывает общую долю правильно классифицированных объектов среди всех объектов выборки.

2. *Error rate* (доля ошибок):

$$\text{Error_rate} = \frac{FP + FN}{TP + TN + FP + FN} = 1 - \text{Accuracy}. \quad (2)$$

Метрика *error rate* показывает долю ошибок бинарной классификации, то есть сумму ошибок 1-го и 2-го рода, деленную на общее количество объектов в выборке.

Accuracy и *error rate* являются простыми интерпретируемыми метриками, однако они не подходят для оценки классификаторов, обученных на несбалансированных выборках². Например, если нам необходимо построить модель для выявления мошеннических кредитных заявок, доля которых составляет 1% от всех заявок, тогда простой классификатор, относящий все кредитные заявки в класс «немошеннические», будет иметь *accuracy* = 99% и *error_rate* = 1%, что является хорошими показателями качества. Однако такой клас-

Accuracy и error rate являются простыми интерпретируемыми метриками, однако они не подходят для оценки классификаторов, обученных на несбалансированных выборках.

¹ Здесь мы используем перевод термина accuracy, взятый из книги А. Мюллера и С. Гвидо «Введение в машинное обучение с помощью Python», так как альтернативный перевод «точность» используется для другой метрики — precision.

² He H., Garcia E.A. Learning from Imbalanced Data. IEEE Trans. Knowledge and Data Engineering. 2009. Vol. 21. Issue 9. P. 1263-1284.

Сергей АФАНАСЬЕВ Анастасия СМИРНОВА

сификатор будет бесполезен для бизнеса, поскольку он не выявляет ни одной мошеннической заявки.

3. *Recall* (полнота), *sensitivity* (чувствительность), *true positive rate* (доля истинно положительных объектов):

$$\text{Recall} = \text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (3)$$

Метрика *recall* (или *sensitivity*, или *TPR*) имеет простую интерпретацию: она показывает долю объектов положительного класса из общей выборки, которые были определены моделью как положительные. Например, если *recall* скоринговой модели составляет 70%, то это значит, что модель смогла выделить (отказать) 70% всех дефолтных заемщиков.

4. *Specificity* (специфичность):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \quad (4)$$

Метрика *specificity* показывает, какую долю от всех отрицательных объектов классификатор выделил как отрицательные. То есть если *specificity* для скоринговой модели составляет 40% — это значит, что модель одобрила 40% всех недефолтных клиентов.

5. *False positive rate* (доля ложно положительных объектов):

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} = 1 - \text{Specificity}. \quad (5)$$

Метрика *FPR* показывает, какая доля отрицательных объектов была ошибочно отнесена к положительным. Например, если *FPR* скоринговой модели составляет 60% — это значит, что модель отказала 60% недефолтных клиентов.

6. *Precision* (точность), *positive predictive value* (прогностическая ценность положительного результата):

$$\text{Precision} = \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (6)$$

Метрика *precision* показывает, какая доля положительных объектов, предсказанных моделью, действительно являются положительными.

Если, например, для кредитного скоринга *precision* составляет 30%, то это значит, что 30% всех отказных заявок действительно были дефолтными, а остальные 70% отказных заявок были недефолтными.

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

7. F_1 -score (F -мера¹):

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (7)$$

F -мера, рассчитанная как среднее гармоническое точности и полноты, является популярной метрикой в задачах с несбалансированными классами. Помимо классической метрики F_1 -score на практике также используют F_β -меру с коэффициентами $\beta = 0,5$ и $\beta = 2$ (табл. 1). В общем случае формула для F_β -меры имеет вид:

$$F_\beta = (\beta^2 + 1) \times \frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}. \quad (8)$$

8. Matthews correlation coefficient (коэффициент корреляции Мэтьюса):

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \quad (9)$$

Коэффициент корреляции Мэтьюса учитывает все переменные матрицы ошибок и устойчив к смене меток классов².

9. LIFT (прирост концентрации³):

$$\text{LIFT} = \frac{\text{Precision}}{(\text{TP} + \text{FN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP})}. \quad (10)$$

Метрика *LIFT* показывает, насколько модель улучшает долю положительных объектов в подмножестве относительно доли в случайно выбранном подмножестве такого же размера.

Формулы для перечисленных и других популярных пороговых метрик представлены в табл. 1.

Таблица 1

Пороговые метрики качества*

Метрика	Формула
ACC	$(\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP})$
ERR	$(\text{FP} + \text{FN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP})$
PPCR	$(\text{TP} + \text{FP}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP})$
TNR	$\text{TN} / (\text{TN} + \text{FP})$

¹ <https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf>.

² <https://clevertap.com/blog/the-best-metric-to-measure-accuracy-of-classification-models/>.

³ https://github.com/esokolov/ml-course-msu/blob/master/ML15/lecture-notes/Sem05_metrics.pdf.

Сергей АФАНАСЬЕВ
Анастасия СМИРНОВА

Окончание табл. 1

Метрика	Формула
REC, SN, TPR	$TP / (TP + FN) = 1 - FNR$
bACC	$0,5 (TNR + TPR)$
SP	$TN / (TN + FP) = 1 - FPR$
FPR	$FP / (TN + FP) = 1 - SP$
FNR	$FN / (TP + FN) = 1 - SN$
LRP	$SN / (1 - SP) = (1 - FNR) / FPR$
LRN	$(1 - SN) / SP = FNR / (1 - FPR)$
PREC, PPV	$TP / (TP + FP)$
FDR	$FP / (TP + FP) = 1 - PPV$
NPV	$TN / (TN + FN)$
FOR	$FN / (TN + FN) = 1 - NPV$
$F_{0,5}$	$1,25 \times PREC \times REC / (0,25 \times PREC + REC)$
F_1	$2 \times PREC \times REC / (PREC + REC)$
F_2	$5 \times PREC \times REC / (4 \times PREC + REC)$
MCC	$(TP \times TN - FP \times FN) / ((TP + FP)(TP + FN)(TN + FP)(TN + FN))^{1/2}$
LIFT	$PREC / (TP + FN) / (TP + TN + FN + FP)$

* ACC – accuracy; ERR – error rate; PPCR – predicted positive condition rate; TNR – true negative rate; REC – recall; SN – sensitivity; TPR – true positive rate; bACC – balanced accuracy; SP – specificity; FPR – false positive rate; FNR – false negative rate; LRP – likelihood ratio positive; LRN – likelihood ratio negative; PREC – precision; PPV – positive predictive value; FDR – false discovery rate; NPV – negative predictive value; FOR – false omission rate; F – F -score; MCC – Matthews correlation coefficient; LIFT – concentration increase; TP – true positives; TN – true negatives; FP – false positives; FN – false negatives.

Не зависящие от порога метрики качества

1. AUC-ROC (площадь под ROC-кривой).

ROC-кривая (receiver operating characteristic curve) описывает множество точек в двумерном пространстве, одна из координат которого соответствует доле ложно положительных объектов ($FPR = 1 - \text{Specificity}$), а другая — доле истинно положительных объектов ($TPR = \text{Sensitivity} = \text{Recall}$).

Чтобы построить ROC-кривую, необходимо упорядочить объекты по убыванию ответов классификатора (вероятностям) и для каждого порогового значения вычислить пары (FPR , TPR). Всего таких пар точек будет $N + 1$, где N — количество объектов в выборке, на которых вычислялся прогноз. Максимальный порог даст классификатор с $TPR = 0$, $FPR = 0$, а минимальный даст классификатор с $TPR = 1$ и $FPR = 1$. Таким образом, ROC-кривая — это кривая с концами в точ-

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

ках (0; 0) и (1; 1), которая последовательно соединяет пороговые точки, упорядоченные по убыванию прогнозов¹.

Площадь под ROC-кривой называется метрикой *AUC-ROC* (Area Under Curve ROC), которая принимает значения от 0 до 1. Если классификатор не допускает ошибок, то *AUC-ROC* будет равна 1. Если классификатор ранжирует объекты случайным образом, то *AUC-ROC* будет близка к 0,5. Соответственно у наихудшего классификатора *AUC-ROC* будет равна 0.

Если *AUC-ROC* принимает значения меньше 0,5, то ROC-кривая лежит ниже базовой линии (диагонали), а классификатор работает хуже случайного. В этой ситуации можно поменять предсказанные метки классов и получить классификатор лучше случайного, то есть ROC-кривая преобразованного классификатора будет лежать выше диагонали. Поэтому для удобства далее будем считать, что ROC-кривая всегда лежит выше диагонали или совпадает с ней, тогда *AUC-ROC* будет принимать значения в диапазоне от 0,5 до 1.

Показатель *AUC-ROC* имеет вероятностную интерпретацию: это вероятность того, что случайно выбранный объект положительного класса имеет оценку принадлежности к положительному классу выше, чем случайно взятый объект отрицательного класса².

2. *Gini index* (индекс Джини).

В банковском моделировании вместо *AUC-ROC* часто используют индекс Джини³, который линейно связан с метрикой *AUC-ROC*:

$$\text{Gini} = 2 \times \text{AUC} - 1. \quad (11)$$

Если ROC-кривая лежит *не ниже* базовой линии (диагонали), то несложно доказать, что индекс Джини равен отношению площади между ROC-кривой и диагональю ко всей площади треугольника над диагональю, которая равна 0,5 (рис. 2б):

$$\text{Gini} = 2 \times \text{AUC} - 1 = 2 \times (\text{AUC} - 0,5) = \frac{\text{AUC} - 0,5}{0,5} = \frac{S_A}{S_A + S_B}. \quad (12)$$

Таким образом, индекс Джини является нормированной метрикой и принимает значения от 0 до 1 (0 — случайный классификатор, 1 — идеальный классификатор).

Показатель *AUC-ROC* имеет вероятностную интерпретацию: это вероятность того, что случайно выбранный объект положительного класса имеет оценку принадлежности к положительному классу выше, чем случайно взятый объект отрицательного класса.

¹ https://github.com/esokolov/ml-course-msu/blob/master/ML15/lecture-notes/Sem05_metrics.pdf.

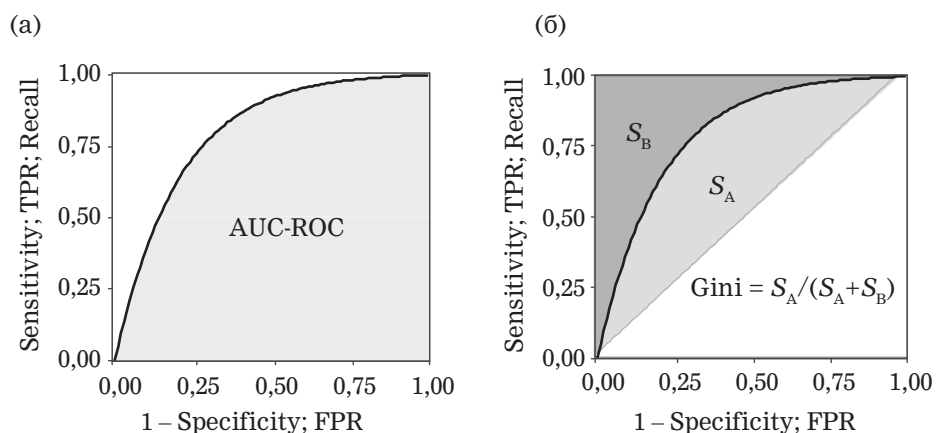
² <https://dyakonov.org/2017/07/28/>.

³ Индекс Джини в машинном обучении не равен экономическому коэффициенту Джини, отражающему степень расслоения общества относительно какого-либо показателя (доход, имущество, недвижимость и т.п.). Общим у этих двух коэффициентов является способ их геометрического вычисления через кривые ROC и Лоуренса соответственно.

Сергей АФАНАСЬЕВ
Анастасия СМЕРНОВА

Рисунок 2

ROC-кривая и AUC-ROC (а), геометрическая интерпретация индекса Джини (б)



Метрика AUC-PR не подходит для сравнения классификаторов, построенных на выборках с разным дисбалансом (в отличие от AUC-ROC, где базовая линия не зависит от дисбаланса).

3. *AUC-PR* (площадь под PR-кривой), *average precision* (средняя точность).

PR-кривая (precision recall curve) строится аналогично ROC-кривой, только по осям абсцисс и ординат откладываются не FPR и TPR, а *recall* (полнота) и *precision* (точность) соответственно (рис. 3а). Напомним, что $\text{recall} = \text{TPR} = \text{Sensitivity}$, то есть ось абсцисс PR-кривой эквивалентна оси ординат ROC-кривой. Площадь под PR-кривой называется метрикой *AUC-PR* (Area Under Curve PR) или метрикой *average precision* (AP)¹.

AUC-PR, как и *AUC-ROC*, может принимать теоретические значения от 0 до 1. Однако у метрики *AUC-PR* есть важное отличие от *AUC-ROC*. Как уже было отмечено, базовая линия ROC-кривой для любого классификатора лежит на отрезке прямой с концами (0; 0) и (1; 1), то есть является диагональю квадрата координатной плоскости. Базовая линия соответствует бесполезному классификатору, который ранжирует объекты случайным образом. Следовательно, *AUC-ROC* для любого бесполезного классификатора всегда будет равна 0,5.

Для PR-кривой базовая линия также имеет форму прямой, однако положение этой линии зависит от дисбаланса выборки: базовая линия PR-кривой проходит через точки (0; d) и (1; d), где d — это

¹ http://www.machinelearning.ru/wiki/images/1/1c/Sem06_metrics.pdf.

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

доля объектов меньшего класса в выборке (дисбаланс)¹. То есть базовые линии PR-кривых у разных классификаторов не обязаны совпадать.

Если считать, что классификаторы всегда являются полезными (т.е. PR-кривые лежат выше базовой линии), то *AUC-PR* будет принимать значения от d до 1. Из этого можно сделать вывод, что метрика *AUC-PR* не подходит для сравнения классификаторов, построенных на выборках с разным дисбалансом (в отличие от *AUC-ROC*, где базовая линия не зависит от дисбаланса). Устранить эту проблему можно с помощью нормированной метрики *AP*.

4. *Normalized average precision* (нормированная средняя точность).

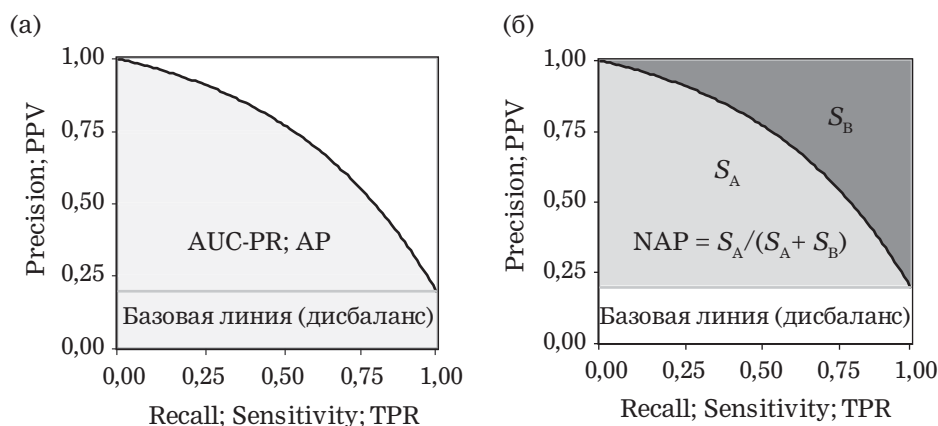
По аналогии с Gini для PR-кривой можно построить нормированную метрику, равную отношению площади между PR-кривой и базовой линией к площади всего прямоугольника, лежащего выше базовой линии (рис. 3б):

$$AP_{\text{normalized}} = \frac{S_A}{S_A + S_B} = \frac{AP - d}{1 - d} = \frac{1}{1 - d} (AP - d) = \frac{1}{1 - d} AP - \frac{d}{1 - d}. \quad (13)$$

Можно увидеть, что при $d = 0,5$ формула (13) принимает вид, аналогичный формуле (11).

Рисунок 3

PR-кривая и AUC-PR (AP) (а), нормированная средняя точность (NAP) (б)



¹ Это утверждение легко доказывается. Беспольный классификатор расставляет метки 0 и 1 случайным образом. Поэтому соотношение $TP / (TP + FP) = \text{Precision}$ будет равно дисбалансу d , а соотношение меток классификатора при этом может быть любым, то есть recall принимает значения от 0 до 1.

Сергей АФАНАСЬЕВ
Анастасия СМИРНОВА

5. *KS* (статистика Колмогорова–Смирнова).

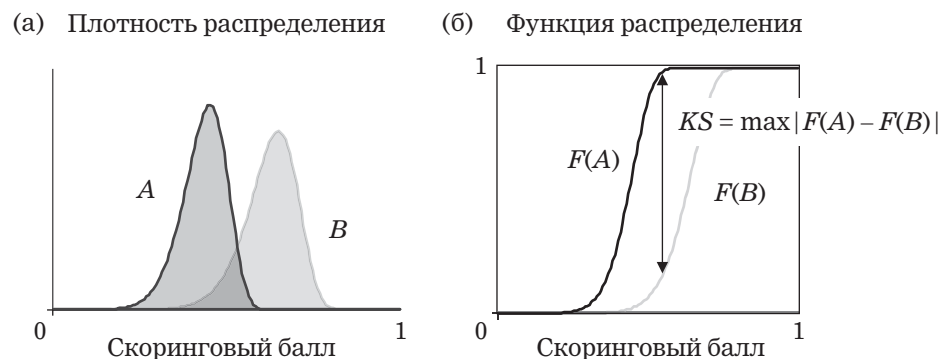
Это еще одна популярная метрика в банковском моделировании, которая определяется как максимальная разность между функциями распределения положительных и отрицательных объектов выборки (рис. 4):

$$KS = \max_i |F(a_i) - F(b_i)|, \quad (14)$$

где $\{a_i\}$ и $\{b_i\}$ — распределения положительных и отрицательных объектов по значениям прогнозов, присвоенных классификатором. Индекс i принимает целочисленные значения от 1 до N , где N — количество объектов в выборке.

Рисунок 4

Статистика Колмогорова–Смирнова



6. *S*-статистика (статистика площади).

Если хотя бы одно из распределений является мультимодальным (имеет несколько локальных мод), то метрика *KS* может принимать заниженные значения. В этих случаях лучше использовать нечувствительные к мультимодальности метрики, такие как *S*-статистика, хи-квадрат, дивергенция Кульбака–Лейблера и др.

S-статистика вычисляется как сумма модулей разностей двух распределений:

$$S = \sum_{i=1}^N \frac{|a_i - b_i|}{2}. \quad (15)$$

S-статистика имеет простую геометрическую интерпретацию: это нормированная сумма площадей под непересекаемыми областями двух распределений. Значения *S*-статистики нормированы и лежат в диапазоне $[0; 1]$: при $S = 0$ распределения полностью совпа-

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

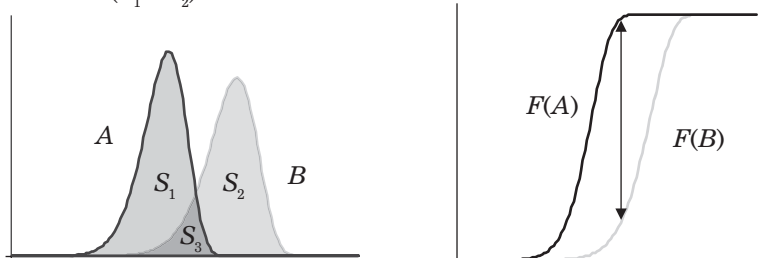
дают (бесполезный классификатор), при $S = 1$ распределения полностью не пересекаются (идеальный классификатор).

Математически можно доказать, что *S-статистика* является обобщением статистики Колмогорова–Смирнова: в случае мономодальных распределений часто будет выполняться равенство $S = KS$, в случае мультимодальных распределений чаще будет выполняться неравенство $S > KS$ (рис. 5)¹.

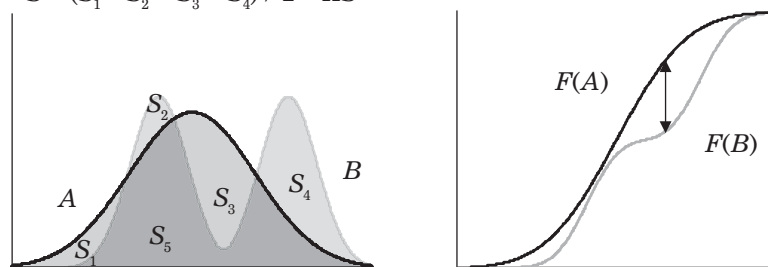
Рисунок 5

Сравнение метрик S и KS

(а) $S = (S_1 + S_2) / 2 = KS$



(б) $S = (S_1 + S_2 + S_3 + S_4) / 2 > KS$



Статистики KS и S относятся к метрикам расстояния между двумя распределениями (дивергенциям). В математической статистике существует множество других метрик расстояния, таких как χ^2 , t -test, MAD -test, статистика Андерсона–Дарлинга и др. В нейросетевом моделировании для оценки отличия классов часто используют дивергенцию Кульбака–Лейблера и дивергенцию Джессена–Шеннона (табл. 2).

¹ Здесь слово «часто» означает, что на практике это будет выполняться для большинства задач (но не всегда). Математическое доказательство этого свойства основано на смене знаков разностей значений двух распределений.

Сергей АФАНАСЬЕВ
Анастасия СМЕРНОВА

Таблица 2

Метрики качества, не зависящие от порога*

Метрика	Анализируемая кривая	Методика вычисления
AUC-ROC	ROC-кривая	Площадь под ROC-кривой
Gini	ROC-кривая	$Gini = 2 \times AUC_ROC - 1$
AUC-CROC	CROC-кривая ¹	Площадь под CROC-кривой
AUC-PR, AP	PR-кривая	Площадь под PR-кривой
AUC LIFT	LIFT-кривая ²	Площадь под LIFT-кривой
NAP	PR-кривая	$NAP = AP / (1 - d) - d / (1 - d)$
KS	Функции распределения	$KS = \max F(a_i) - F(b_i) $
S-test	Плотности распределения	$S = \sum a_i - b_i / 2$
Chi	Плотности распределения	$Chi = \sum ((a_i - b_i) / (0,5)) / (a_i + b_i)$
T-test	Плотности распределения	$T = (E(a_i) - E(b_i)) / ((\text{var}(a_i) / N_a) + (\text{var}(b_i) / N_b))^{(0,5)}$
MAD-test ³	Плотности распределения	$MAD = \sum a_i - b_i / K, 1 \leq i \leq K$
AD-test ⁴	Плотности распределения	$AD = (\sum (N_i \times Z(N_a + N_b - N_a \times i))^{(0,5)}) / (i \times Z(N_a + N_b - i)) / (N_a \times N_b), 1 \leq i \leq N_a + N_b$
KLD	Плотности распределения	$KLD(a_i b_i) = \sum a_i \times \log(a_i / b_i)$
JSD	Плотности распределения	$JSD(a_i b_i) = (KLD(a_i (a_i + b_i) / 2) + KLD(b_i (a_i + b_i) / 2)) / 2$

* AUC-ROC – area under curve ROC; Gini – Gini index; AUC-CROC – area under curve concentrated ROC; AUC-PR – area under curve PR; AP – average precision; AUC LIFT – area under curve LIFT; NAP – normalized average precision; KS – Kolmogorov-Smirnov test; Chi – Pearson's chi-squared test; T-test – Welch's t-test, unequal variances t-test; MAD – mean absolute deviation; AD-test – Anderson-Darling test; KLD – Kullback-Leibler divergence; JSD – Jensen-Shannon divergence.

Недостатки AUC-ROC и Gini для несбалансированных выборок

Разные метрики качества имеют свои достоинства и недостатки. Как было показано ранее, метрика ассурасу плохо работает на несбалансированных выборках, когда объектов одного класса значительно меньше, чем объектов другого. Такая же проблема возникает при использовании метрик *AUC-ROC* и *Gini*. Чтобы продемонстрировать это, рассмотрим задачу выявления мошеннических транзакций из

¹ CROC-кривая строится через преобразование осей ROC-кривой и выполняет функцию «лупы» для одного из участков ROC-кривой (обычно раннего участка с низкими значениями FPR).

² <http://mrvar.fdv.si/pub/mz/mz3.1/vuk.pdf>.

³ <https://arxiv.org/pdf/1311.4787.pdf>.

⁴ http://www.jaqm.ro/issues/volume-6,issue-3/pdfs/1_engmann_cousineau.pdf.

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

множества транзакций по банковским картам. Допустим, что всего имеется 1 000 100 транзакций, из которых 100 являются мошенническими.

Если нам удастся построить идеальный классификатор, то его *TPR* (*sensitivity, recall*) будет равна единице, а *FPR* будет равна нулю (*AUC-ROC* и *Gini* для такого алгоритма будут равны 1). Рассмотрим теперь плохой классификатор, дающий положительный ответ на 95 мошеннических транзакций из 50 000 хороших. Такой алгоритм скорее будет бесполезным, но при этом он имеет $TPR = 0,95$ и $FPR = 0,05$, что очень близко к идеальным показателям. Таким образом, если выборка несбалансированная, то *AUC-ROC* и *Gini* могут давать неадекватную оценку качества работы классификатора, поскольку измеряют долю неверно принятых объектов относительно общего числа отрицательных. Например, если классификатор присваивает мошенническим транзакциям скоринговый балл с 50 001 по 50 101, то он будет иметь $AUC-ROC = 0,95$ и $Gini = 0,90$: это очень высокие показатели качества, хотя сам классификатор является скорее бесполезным¹.

Если перейти на оценку качества алгоритма через *PR-кривую*, то классификатор, присваивающий мошенническим транзакциям скоринговые баллы с 50 001 по 50 101, будет иметь $AUC-PR = 0,001$ и $NAP = 0,0009$: это более адекватные показатели качества для данного классификатора.

В качестве другого примера можно рассмотреть ROC-кривые для двух классификаторов, которые симметричны относительно диагонали, проходящей через точки (0; 1) и (1; 0). Для этих ROC-кривых площади под графиками будут равны, то есть будут равны метрики *AUC-ROC*, а также индексы *Gini*. Таким образом, можно ошибочно сделать вывод, что классификаторы одинаковые. Однако если выбирать порог, то может оказаться, что при равных *FPR* показатель *TPR* будет отличаться, например, в 2 раза (рис. 6а). Аналогичную проблему можно увидеть, если зафиксировать *TPR* (рис. 6б).

Приведенные примеры демонстрируют недостатки ROC-кривой и преимущества PR-кривой на несбалансированных выборках. Однако специально подобранные примеры не обладают обобщающей способностью и не доказывают, что ROC-кривая работает хуже PR-кривой на случайной несбалансированной выборке. Общее доказательство

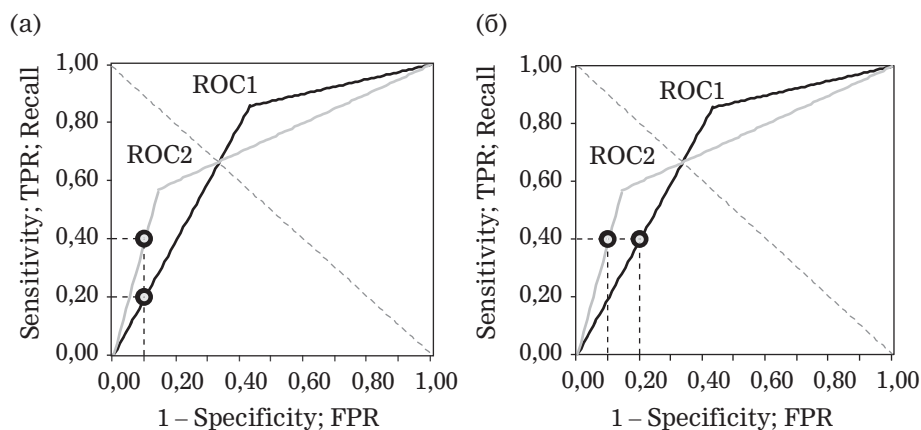
Практически все банки используют *Gini* для оценки качества классификатора. Тем не менее, такие метрики, как *Gini* и *AUC-ROC*, плохо работают на несбалансированных выборках, а в банках почти все выборки являются несбалансированными.

¹ За основу мы взяли пример из лекции Евгения Соколова о ранжировании поисковых выдач: http://www.machinelearning.ru/wiki/images/1/1c/Sem06_metrics.pdf.

Сергей АФАНАСЬЕВ
Анастасия СМИРНОВА

Рисунок 6

**Одинаковые значения AUC-ROC могут давать
различный результат при выборе порога: $FPR_1 = FPR_2$,
 $TPR_1 = 0,5 \times TPR_2$ (а); $FPR_1 = 2 \times FPR_2$, $TPR_1 = TPR_2$ (б)**



этого эффекта приводят в своем исследовании норвежские ученые Takaya Saito и Mark Rehmsmeier¹.

В первой части исследования они сгенерировали выборки с различными распределениями скоринговых баллов, то есть рассмотрели различные по качеству классификаторы. С этой целью они сгенерировали выборки для пяти типов классификаторов:

- 1) случайный (random);
- 2) плохой (poor early retrieval);
- 3) хороший (good early retrieval);
- 4) отличный (excellent);
- 5) идеальный (perfect).

Далее для двух типов выборок, сбалансированной и несбалансированной, были построены четыре типа кривых для оценки качества классификации: ROC-кривые, CROC-кривые, Cost Curves (CC)² и PR-кривые. Оказалось, что кривые ROC, CROC и CC показывают одинаковое качество классификаторов на сбалансированной и несбалансированной выборках. И только PR-кривые показывали отличия и, в частности, плохое качество классификаторов на несбалансированной выборке.

¹ Saito T., Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE 10(3):e0118432, March 2015. DOI: 10.1371/journal.pone.0118432.

² Drummond C., Holte R.C. Cost curves: An improved method for visualizing classifier performance. Springer Science + Business Media, LLC 2006.

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

Во второй части исследования были проанализированы 58 медицинских статей, опубликованных с 2002 по 2012 гг. на научном ресурсе PubMed. Оказалось, что в 66,7% исследований, проведенных на несбалансированных выборках, для оценки качества классификаторов использовались ROC-кривая и метрика *AUC-ROC*.

И наконец, в третьей части исследования ученые взяли одну из анализируемых статей с PubMed и построили классификатор на опубликованных данных этого исследования. После этого они сравнили метрики качества классификатора *AUC-ROC* и *AUC-PR*. Оказалось, что *AUC-ROC* показала хорошее качество алгоритма, как это было продемонстрировано в оригинальном исследовании. Однако показатель *AUC-PR* оказался крайне низким для того, чтобы считать классификатор хорошим, как утверждалось в оригинальном исследовании.

Эти результаты не просто настораживают, а ставят под сомнение результаты 2/3 всех медицинских исследований всего лишь из-за одной неправильно выбранной метрики. Но как это относится к банковской сфере и, в частности, к банковскому моделированию?

Во-первых, в банковском моделировании выборки практически всегда являются несбалансированными. В классических задачах скоринга доля дефолтных заемщиков не будет превышать 10–15% (если это не МФО с высоким уровнем просрочки). В коллекшн-скоринге (recovery rate модели) процент дефолтных заемщиков, полностью погасивших кредит, также крайне мал. В антифрод-моделировании доля мошеннических кредитов еще меньше — от десятых долей до пары процентов, а доля транзакционного мошенничества измеряется сотыми и даже тысячными долями процента.

Во-вторых, практически все банки используют *Gini* для оценки качества. Например, на четырех последних конференциях по скорингу, проводимых ИД «Регламент»¹, почти во всех презентациях докладчиков *Gini* была основной метрикой для демонстрации качества моделей. Лишь в одном докладе показывали PR-кривую.

Возникает вопрос: если *AUC-ROC* и *Gini* — плохие метрики для банковских задач, то какие метрики необходимо использовать?

Прежде чем ответить на этот вопрос, рассмотрим еще две задачи из реальной практики, демонстрирующие недостатки метрик *AUC-ROC* и *Gini*.

¹ <http://scorconf.ru>.

Сергей АФАНАСЬЕВ
Анастасия СМИРНОВА

Эксперименты

1. Сравнение алгоритмов ML для прогнозирования recovery rate

Первая задача заключалась в построении recovery rate модели для подразделения взыскания просроченной задолженности: спрогнозировать, что клиент, допустивший просрочку более X дней, погасит более Y% суммы долга.

Цель задачи: сравнить разные ML-методы для прогнозирования recovery rate.

Сравниваемые алгоритмы: Logistic Regression, Neural Networks (fully connected).

Исходная выборка:

- размер выборки: 922 150 записей;
- класс 1 (good): 11 241 запись (1,2% выборки);
- класс 2 (bad): 910 909 записей (98,8% выборки);
- train sample: 737 720 записей;
- test sample: 184 430 записей;
- количество признаков: 958.

Pipeline:

1) подготовка выборки: удаление некорректных данных, обработка пропусков, биннинг;

2) отбор признаков: с помощью XGBoost отобрано 215 значимых переменных (применялись нормализация и подбор гиперпараметров по кросс-валидации);

3) обучение моделей: нормализация признаков, подбор гиперпараметров по кросс-валидации, сравнение качества на тестовой выборке;

4) сравнение качества построенных моделей.

Для подбора гиперпараметров использовалась метрика *AUC-ROC* — это было обязательным требованием в задаче, так как классификаторы сравнивались с ранее построенными моделями, которые оценивались по метрике *AUC-ROC*.

Нейронные сети обучались на двух архитектурах с подбором гиперпараметров:

- 1) полносвязная сеть с двумя скрытыми слоями;
- 2) полносвязная сеть с пятью скрытыми слоями.

Кроме подбора гиперпараметров, в нейронных сетях использовались ранняя остановка и фиксированные гиперпараметры: BATCH_SIZE = 128; DROPOUT = 0,2; OPTIMIZER = SGD(); Activation (hidden): RELU; Activation (out): softmax.

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

Результаты обученных моделей с гиперпараметрами представлены в табл. 3.

Таблица 3

Сравнение моделей*

# Model	Гиперпараметры					Метрики качества, %					
	HL	N-HL	Reg_C	Epoch	Class_w	Precision	Recall	F ₁ -score	AUC-ROC	Gini	AUC-PR
1. NN_H5_N200	5	200		120	0,030	8,3	67,1	14,8	88,9	77,8	17,6
2. NN_H2_N400	2	400		89	0,014	5,7	78,6	10,6	88,8	77,6	16,8
3. Log_reg			0,01		0,040	11,1	50,6	18,2	88,1	76,3	15,8
4. NN_H2_N400	2	400		73	0,050	11,3	56,3	18,8	88,9	77,7	17,8
5. NN_H5_N300	5	300		92	0,050	11,5	54,2	19,0	88,7	77,4	17,8

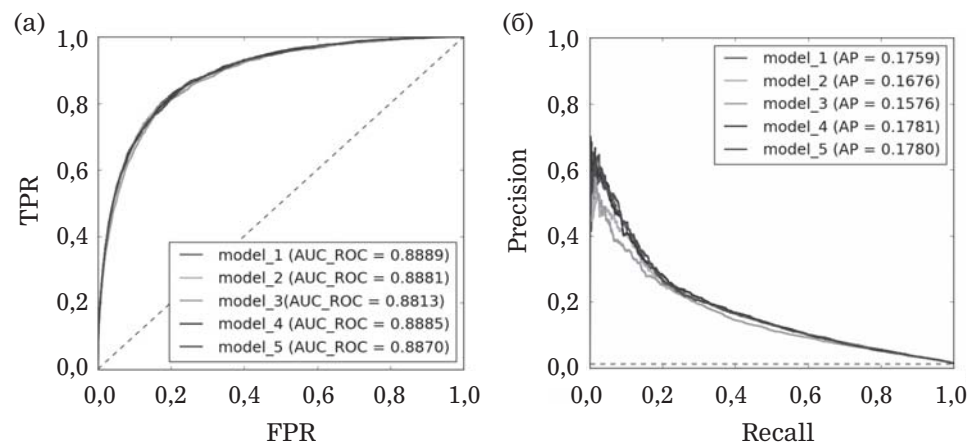
* NN_H5_N200 — полносвязная нейронная сеть с пятью скрытыми слоями, по 200 нейронов в каждом скрытом слое; NN_H2_N400 — полносвязная нейронная сеть с двумя скрытыми слоями, по 400 нейронов в каждом скрытом слое; Log_reg — логистическая регрессия; NN_H5_N300 — полносвязная нейронная сеть с пятью скрытыми слоями, по 300 нейронов в каждом скрытом слое.

Первые три модели, представленные в табл. 3, обучены на лучших гиперпараметрах, подобранных по кросс-валидации (метрика — *AUC-ROC*). Видно, что нейронные сети показывают качество по *AUC-ROC* и *Gini* чуть выше логистической регрессии. При этом разница *Gini* между лучшей нейронной сетью и логистической регрессией составляет всего 1,5%. Однако если смотреть на показатели *precision* и *recall*, то разница уже существеннее, при этом *precision* выше у логистической регрессии, а *recall* — у нейронных сетей. По этим значениям метрик сложно сделать вывод, какая модель лучше. Поэтому для нейронных сетей мы выбрали другие гиперпараметры, которые показали значение *precision*, близкое к логистической регрессии. Потом мы сравнили качество моделей 3, 4 и 5 и увидели, что разница в *Gini* оказалась несущественной, так же как и в предыдущих сравнениях. Разница *precision* была незначительной по условиям подбора (~0,4%). А вот показатели *recall* отличались на 4,4%, что можно интерпретировать следующим образом: построенная нейронная сеть выделяет клиентов из положительного класса на 4,4% больше, чем логистическая регрессия (при прочих равных параметрах). Кроме того, по метрикам *AUC-PR* (*AP*) и *F₁-score* видно, что все построенные модели не обладают высоким качеством, как это ошибочно можно предположить, если оценивать модели по *AUC-ROC* или *Gini* (рис. 7).

Сергей АФАНАСЬЕВ
Анастасия СМЕРНОВА

Рисунок 7

Высокая AUC-ROC, низкая AUC-PR



2. Проверка переобучения моделей сотовых операторов

Вторая задача заключалась в проверке комбинированной скоринговой карты, построенной на трех скоринговых баллах:

- 1) банковском скорбалле;
- 2) скорбалле на данных социальных сетей;
- 3) скорбалле на данных сотовых операторов.

Парные комбинированные скоркарты «банк + соцсети» и «банк + операторы» давали одинаковое качество по метрике *Gini*. Необходимо было проверить — дает ли второй внешний источник данных дополнительную прибавку к качеству модели, то есть измерить качество тройной комбинированной скоркарты «банк + соцсети + операторы».

Проблема заключалась в том, что тестовые выборки, на которых проверялось качество парных комбинированных скорбаллов, не пересекались. То есть соцсети тестировались на одной выборке, а операторы — на другой. Однако часть тестовой выборки для соцсетей была использована как обучающая выборка для кастомизации скорбалла операторов. Кроме того, была информация, что по модели операторов значение *Gini* на обучающей выборке не сильно отличается от значения *Gini* на тестовой выборке, то есть модель не переобучена. При таких условиях можно было использовать пересечение обучающей выборки операторов с тестовой выборкой соцсетей и на этих скорбаллах построить тройную комбинированную скоркарту.

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

Однако надо было проверить, действительно ли скорбаллы операторов, полученные на обучающей выборке, были не переобучены.

Цель задачи: оценить степень переобучения модели операторов, сравнив результаты на обучающей и тестовой выборках.

Исходные данные:

- train sample: 94 700 скорбаллов (18 месяцев);
- out-off-time test sample: 14 600 скорбаллов (2 месяца).

Созданные выборки:

- sample_1 (train): 75% train; 25% test. Вся train-выборка: 94 700 скорбаллов;
- sample_2 (out_off_time): 66,7% train; 33,3% test. Out-off-time test-выборка: 14 600 скорбаллов.

Алгоритм: логистическая регрессия.

Переменные:

- независимые: банковский скорбалл; скорбалл операторов;
- целевая: 90-дневная просрочка на N -платеже.

Так как выборки были несбалансированные, для подбора гиперпараметров мы использовали метрику F_1 -score. Гиперпараметры отбирались по кросс-валидации (5-fold). Для обеих выборок были подобраны оптимальные гиперпараметры регуляризации и балансировки весов: $Reg_C = 0,1$, $cl_w = 0,05$. Качество финальных моделей проверили и сравнили на тестовых выборках (test sample_1 и test sample_2). Полученные результаты представлены в табл. 4.

По метрике AUC -ROC на тестовых выборках видно, что все модели имеют практически одинаковое качество с разницей AUC -ROC 0,0132, что в относительном выражении составило 2,02%. Можно сделать вывод, что модели на обучающей выборке не сильно переобучены. Однако если смотреть на метрики $precision$, $recall$, F_1 -score и NAP , то разница уже существенна: показатели $recall$ на тестовой «train-выборке» (sample_1) и тестовой «out-off-time выборке» (sample_2)

Таблица 4

Метрики качества моделей, построенных на sample_1 и sample_2

	AUC-ROC	Precision	Recall	F_1 -score	NAP
Sample_1 train	0,6523	0,0747	0,6614	0,1342	
Sample_2 train	0,6982	0,0687	0,4633	0,1196	
Sample_1 test	0,6525	0,0768	0,6756	0,1379	0,0349
Sample_2 test	0,6657	0,0629	0,3636	0,1072	0,0449
Относительная разница метрик на test-sample: $(S_2 - S_1) / S_1$	2,02%	-18,10%	-46,18%	-22,26%	28,65%

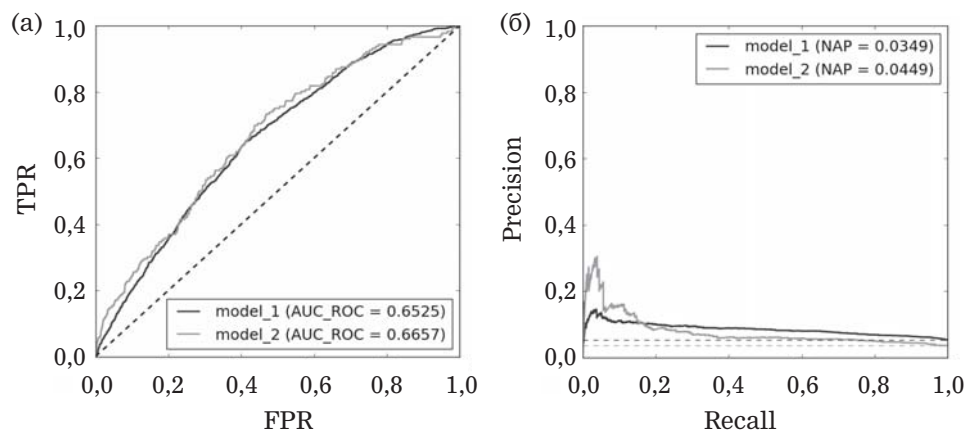
Сергей АФАНАСЬЕВ
Анастасия СМИРНОВА

отличаются на 46,18% в относительном выражении (0,6756 против 0,3636). Показатели *precision* и *F₁-score* также показывают значимое различие. На ранних участках PR-кривая у первой модели показывает качество хуже, чем у второй модели. Однако на поздних участках, наоборот, качество первой модели значительно лучше.

Поскольку модели построены на выборках с разным дисбалансом, для сравнения качества были посчитаны метрики *NAP* (нормированная средняя точность). По PR-кривым и значениям *NAP* видно, что модель на «train-выборке» (sample_1) сильно переобучена, чего не видно по ROC-кривым и *AUC-ROC* (рис. 8). Из этих результатов можно сделать вывод, что скоринговые баллы, построенные на обучающей выборке сотовых операторов, нельзя использовать для обучения тройной комбинированной модели «банк + соцсети + операторы».

Рисунок 8

Переобучение не видно по ROC-кривой и видно по PR-кривой



Бизнес-метрики для банковских моделей

Преимущества PR-кривой над ROC-кривой, казалось бы, закрывают вопрос о выборе метрик для моделей, построенных на несбалансированных выборках. Но так ли все идеально с PR-кривой и метриками *AUC-PR* и *NAP*?

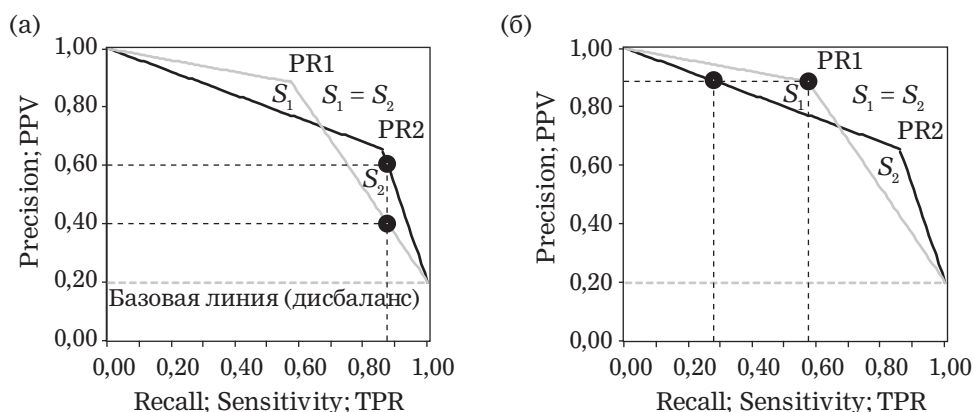
Анализируя различные виды PR-кривой, можно продемонстрировать, что метрики *AUC-PR* и *NAP* также имеют ряд недостатков, связанных с тем, что они являются интегральными (т.е. при вычислении теряют часть информации).

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

На рис. 9 показаны две симметричные PR-кривые, для которых нормированные точности NAP равны. Однако при выборе порога с одинаковыми значениями $recall$ значения $precision$ существенно различаются (рис. 9а). И наоборот, при фиксированном $precision$ сильно различаются значения $recall$ (рис. 9б).

Рисунок 9

Одинаковые значения AUC-PR могут давать различные результаты при выборе порога:
TPR1 = TPR2, Precision1 = (2/3) × Precision2 (а);
TPR1 = 2 × TPR2, Precision1 = Precision2 (б)



Если вернуться к интерпретации метрик $precision$ и $recall$, то при разработке скоринговых моделей могут возникнуть и другие вопросы, например: какое соотношение значений $precision$ и $recall$ будет оптимальным для достижения поставленных бизнес-целей?

Вышеперечисленные проблемы стандартных метрик подводят нас к выводу, что для оценки качества моделей необходимо использовать бизнес-метрику, отражающую уровень достижения конкретной (оцифрованной) бизнес-цели. Например, в задачах скоринга основной бизнес-целью является максимизация прибыли от всего входящего потока кредитных заявок. То есть для скоринговых моделей бизнес-метрика должна отражать прибыль портфеля, которая выражается через показатели «хороших» и «плохих» выданных кредитов. Таким образом, если все одобренные кредиты разделить на «хорошие» (не попавшие в просрочку) и «плохие» (попавшие в просрочку), то прибыль портфеля можно вычислить по формуле:

Сергей АФАНАСЬЕВ
Анастасия СМЕРНОВА

$$P = m \times G - (1 - RR) \times B, \quad (16)$$

где P (Profit) — прибыль портфеля;
 G (Good) — портфель «хороших» кредитов;
 B (Bad) — портфель «плохих» кредитов;
 m (margin) — маржа по «хорошим» кредитам с учетом ОПЕХ-расходов;
 RR (recovery rate) — коэффициент восстановления «плохих» кредитов (какая часть просроченных кредитов взыскана с должников).

В обозначениях матрицы ошибок формула (16) переписывается в виде:

$$P = m \times TN \times s - (1 - RR) \times FN \times s, \quad (17)$$

где TN — количество истинно отрицательных объектов («хорошие» кредиты);
 FN — количество ложно отрицательных объектов («плохие» кредиты);
 s — средний чек по выданным кредитам.

В портфельном риск-менеджменте вместо маржи m обычно используют нулевой таргет t_0 . Нулевой таргет показывает уровень просрочки, при котором портфель дает нулевую прибыль (точка безубыточности). Нулевые таргеты вычисляются для каждого выбранного платежа с определенным выбранным количеством дней просрочки. Если записать определение нулевого таргета, а также нулевую прибыль через маржинальность, получим систему уравнений:

$$\begin{cases} t_0 = \frac{B_0}{B_0 + G_0} \\ 0 = m \times G_0 - (1 - RR) \times B_0, \end{cases} \quad (18)$$

из которой следует:

$$\begin{cases} \frac{B_0}{G_0} = \frac{t_0}{1 - t_0} \\ m = (1 - RR) \times \frac{t_0}{1 - t_0} \end{cases}. \quad (19)$$

Таким образом, уравнение (17) переписывается через нулевой таргет t_0 в следующем виде:

$$P = (1 - RR) \left(\frac{t_0}{1 - t_0} \times TN - FN \right) \times s. \quad (20)$$

Если необходимо сравнивать разные модели, построенные на разных выборках, то необходимо нормировать метрику P на количество всех поступивших кредитных заявок:

$$P_{\text{norm}} = \frac{(1 - RR) \left(\frac{t_0}{1 - t_0} \times TN - FN \right) \times s}{TP + TN + FN + FP}. \quad (21)$$

Для оценки качества моделей необходимо использовать бизнес-метрику, отражающую уровень достижения конкретной (оцифрованной) бизнес-цели. Например, для скоринговых моделей бизнес-метрика должна отражать прибыль портфеля, которая выражается через показатели «хороших» и «плохих» выданных кредитов.

Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании

Формула (21) показывает, сколько прибыли заработает банк с каждой поданной кредитной заявки (не обязательно одобренной).

Стоит отметить, что показатели $(1 - RR)$ и s в формуле (20) являются константными множителями, то есть можно поделить на них обе части уравнения и максимизировать полученную бизнес-метрику без учета этих констант:

$$\frac{P}{(1 - RR) \times s} = \frac{t_0}{1 - t_0} \times TN - FN. \quad (22)$$

Таким образом, если не стоит задача сравнения качества моделей на разных портфелях, то можно использовать формулу (22), для которой необходима оценка только нулевого таргета t_0 для портфеля.

Аналогичную метрику можно использовать для моделей детектирования заявочного мошенничества, где в качестве целевой переменной выбирается просрочка по группе потенциально мошеннических кредитов. В этом случае можно использовать формулу (17) для вычисления profits P , однако стоит учитывать, что показатель *recovery rate* по мошенническим заявкам будет ниже, чем по общему портфелю, поэтому RR надо вычислять отдельно для потенциально мошеннических кредитов.

Для моделей выявления внутреннего мошенничества лучше использовать другую бизнес-метрику — потенциальный убыток. Данная метрика позволяет настраивать правила на раннее обнаружение мошенничества, а не на констатацию уже свершившегося убытка.

В задачах CRM для оценки моделей отклика клиентов необходимо максимизировать прибыль от выданных кредитов за вычетом расходов на звонки или SMS:

$$P = m \times TP \times s - c \times (TP + FP), \quad (23)$$

где TP — количество истинно положительных объектов (откликнувшиеся клиенты);

FP — количество ложно положительных объектов (не откликнувшиеся клиенты);

m — маржа на выданный кредит;

s — средний чек по выданным кредитам;

c — затраты на звонки (или SMS) в пересчете на одну одобренную заявку (включая невыданные кредиты).

Заключение

На протяжении многих веков древнегреческий философ Аристотель оставался непоколебимым авторитетом в зоологии. Он лично классифицировал порядка 540 разновидностей животных и изучил внут-

Сергей АФАНАСЬЕВ Анастасия СМИРНОВА


ренное строение по меньшей мере 50 видов. Более 2000 лет никто не допускал мысли, что Аристотель мог в чем-то ошибаться. Так, в некоторых источниках пишут, что Аристотель приписывал мухе восемь лап. И только в середине XVIII в. шведский зоолог Карл Линней усомнился в этом утверждении, пересчитал у мухи лапы и выяснил, что их всего шесть.

Считается, что современная наука началась с принципа методологического сомнения, предложенного Декартом во времена, когда лучшим методом научного исследования было цитирование мнений авторитетов типа Аристотеля. Согласно принципу Декарта, идея не становится научной только от того, что ее защищает выдающийся ученый. Во всем необходимо сомневаться и обо всем можно спорить. Так, высказывается мнение, что история про восьминогую муху Аристотеля является чьим-то вымыслом, кочующим из учебника в учебник¹.

К ошибкам, и даже массовым, склонна и современная наука. Норвежские ученые Takaya Saito и Mark Rehmsmeier показали, что в 67% медицинских исследований использовалась неподходящая для оценки моделей метрика AUC-ROC, которая плохо работает на несбалансированных медицинских данных².

Метрики Gini и AUC-ROC применяются в банковском скоринге уже много лет. В этой статье мы продемонстрировали, что данные метрики плохо работают при оценке банковских моделей по все той же причине — несбалансированности банковских выборок. Но, несмотря на то что уже несколько лет публикуются статьи на эту тему, финансовые организации продолжают использовать метрики Gini и AUC-ROC как основные для сравнения качества своих моделей.

Сейчас отрасль Data Science бурно развивается. В машинном обучении предлагается много различных инструментов, которые легко обобщаются и переносятся из одной задачи в другую. И, как мы видим, довольно часто эти инструменты не анализируются на предмет корректности их использования. Но, в отличие от медицины, где на кону стоят человеческие жизни, банки теряют всего лишь часть прибыли.

Мы предлагаем усомниться в приведенном утверждении и проверить предложенные нами бизнес-метрики, которые определяются через бизнес-цели банка и выражаются в денежном эквиваленте. 

¹ <https://astrei.livejournal.com/12741.html>.

² Saito T., Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE 10(3):e0118432, March 2015. DOI: 10.1371/journal.pone.0118432.