DOI 10.37882/2223-2966.2023.12-2.25

УСОВЕРШЕНСТВОВАНИЕ МЕТОДА ОЦЕНИВАНИЯ ПРИБЫЛЬНОСТИ ГАРАНТИЙ ИСПОЛНЕНИЯ КОНТРАКТА НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

ENCHANCING A MACHINE LEARNING-BASED METHOD TO EVALUATE THE PROFITABILITY OF CONTRACT EXECUTION GUARANTEES

P. Protasov
E. Potekhina

Summary. The present study is focused on investigating the intricacies associated with the concurrent utilization of two distinct machine learning models, each developed with different specifications and dedicated to forecasting the likelihood of contract non-performance during the issuance of a bank guarantee. Within the context of this research, a comprehensive approach for evaluating the profitability of a bank guarantee portfolio is proposed, taking into account the magnitude of the commission remuneration linked to the issuance of the guarantee, the expenditure incurred for acquiring the requisite data for model development and replication, and a quantitative assessment of the risk of contract default, under which the bank guarantee is provided. Conclusions are drawn concerning the connection between the scale of data acquisition costs pertinent to model development and replication and the overall yield of the guarantee portfolio. These elucidated methodologies offer a framework for devising strategies to determine the optimal threshold of the model assessment based on the profitability metric, while factoring in the pertinent business metrics that characterize the specific operational context.

Keywords: binary classification, gradient boosting, confusion matrix, profitability.

Протасов Павел Сергеевич
Аспирант, ФГБОУ ВО «Российский Государственный Социальный Университет»
1pcv@mail.ru
Потехина Елена Витальевна
Доктор экономических наук, профессор, ФГБОУ ВО «Российский Государственный Социальный Университет
elengapotechina@mail.ru

Аннотация. Работа посвящена аспектам совместного использования двух моделей машинного обучения, разработанных на разных характеристиках и предсказывающих риск неисполнения контракта, в процессе выдачи банковской гарантии. Предлагается метод оценивания прибыльности портфеля банковских гарантий, учитывающий размер комиссионного вознаграждения за выдачу гарантии, стоимость расходов на данные, необходимые для разработки и воспроизведения модели, а также количественную оценку риска неисполнения контракта, под который предоставляется банковская гарантия. Сделаны выводы относительно взаимосвязи между величиной расходов на данные, необходимые для разработки и воспроизведения модели, и доходностью портфеля гарантий. Описанные подходы позволяют сформулировать стратегии нахождения оптимального уровня отсечения модельной оценки на основе показателя прибыльности, учитывая бизнес-метрики конкретной задачи.

Ключевые слова: бинарная классификация, градиентный бустинг, матрица неточностей, прибыльность.

Введение

Высокие темпы роста рынка государственных закупок РФ в последние годы свидетельствуют об активном участии большого количества игроков на этом рынке, включая не только крупные компании, но и организации среднего и малого бизнеса. Внедрение и дальнейшее развитие цифровых технологий в сфере государственных закупок сделали процесс проведения закупки более эффективным и прозрачным, одновременно создав условия высокой конкуренции. Одним из основных способов обеспечения закупки в соответствии с требованиями Федерального закона «О контрактной системе в сфере закупок товаров, работ, услуг для обеспечения государственных и муниципальных нужд» от 05.04.2013 г. № 44-ФЗ (далее 44-ФЗ) является предоставление обеспечения по государственной закупке в виде банковской гарантии [5]. Банковская гарантия представляет собой такой вид обеспечения, при котором банк гарантирует заказчику выплату суммы обеспечения, в случаях, если подрядчик нарушит, либо не сможет выполнить условия контракта в установленный срок.

Большой объем информации по государственным закупкам 44-ФЗ, публикуемый и доступный в настоящее время на официальном портале «Единая информационная система в сфере закупок» [5], открывает возможности для проведения разносторонних аналитических исследований и разработки моделей машинного обучения на основе этих данных. Исходя из конкретных целей, для

моделирования могут быть использованы и другие виды данные. К таким данным можно отнести финансовую отчетность подрядчика, информацию о его кредитной истории, судебных тяжбах. Эти данные позволят увеличить точность модельной оценки, но некоторые из этих данных могут быть получены только на коммерческой основе (например, запрос кредитной истории подрядчика в бюро кредитной истории).

С ростом доступности источников данных при решении задач прогнозирования рисков, связанных с государственными закупками, все чаще используются подходы, основанные на математическом моделировании и машинном обучении. Подход на основе машинного обучения предложен авторами Кулаевым М.А., Домашовой Д.В. [4], которые в своей статье представили методику решения задач классификации на базе различных модификаций моделей на основе деревьев решений. Авторами Ивановым Н.В., Валпетерс М.Л., Киреевым И.А. [3] было рассмотрено несколько решений задачи определения вероятности исполнения контракта в строительной отрасли на основе регрессионной модели, а также нелинейных алгоритмов (дерево решений, случайный лес). Ещё один подход к оценке рисков государственных закупок, использующий такие алгоритмы машинного обучения, как случайный лес, классический градиентный бустинг и CatBoost, описан авторами Елисеевым Д., Романовым Д. [2]. Описанные методы позволили преодолеть ограничения, связанные с учетом большого количества параметров в модельных оценках, но оставили открытым вопрос выражения стоимости риска неисполнения государственного контракта как основного аспекта при принятии решения о выдачи банковской гарантии.

Постановка задачи и решение

В статье представлен метод оценивания прибыльности портфеля банковских гарантий при условии совместного использования модельных оценок риска неисполнения контракта, полученных по двум моделям машинного обучения (на примере гарантий исполнения государственного контракта 44-ФЗ). Описанный метод реализуется на основе модельных оценок риска и стоимости расходов на получение дополнительных данных, необходимых для обучения и воспроизведения модели.

Моделирование риска неисполнения контракта относится к задачам бинарной классификации, где прогноз события зависит от порогового уровня отсечения модельной оценки, определенного с помощью матрицы неточностей.

Модель

$\hat{Y}=0$ $\hat{Y}=1$

$$M(\tau)=\begin{bmatrix} TN(\tau) & FP(\tau) \\ FN(\tau) & TP(\tau) \end{bmatrix} \begin{matrix} Y=0 \\ Y=1 \end{matrix} \quad \begin{matrix} \text{Реальность} \\ \end{matrix} \qquad (1)$$

где M_τ — матрица неточностей бинарной классификации;

$\hat{Y}$ — прогнозный исход события на основе модельной оценки;

Y — фактический исход события (позитивный исход при $Y=1$; негативный исход $Y=0$);

τ — пороговый уровень отсечения модельной оценки;

$TN(\tau)$ — результат классификации, при котором правильно определен негативный исход;

$FP(\tau)$ — результат классификации, при котором ошибочно определен позитивный исход;

$FN(\tau)$ — результат классификации, при котором ошибочно отвергнут позитивный исход;

$TP(\tau)$ — результат классификации, при котором правильно определен позитивный исход.

Опираясь на информацию, представленную в формуле (1) прибыль по заданному портфелю гарантий в зависимости от уровня отсечения τ можно рассчитать по формуле:

$$P_\tau=\left(\sum_{\substack{i=1,\\ \forall i\in \\ (S_i\le\tau)}}^{n}\left(FN\_FEE_i+TN\_FEE_i-FN\_AMT_i\right)\right)-C \qquad (2)$$

где P_τ — прибыль (в денежных единицах);

S_i — модельная оценка вероятности неисполнения контракта для i-гарантии (в долях единицы);

i — наблюдение (гарантия) в заданной выборке (портфеле);

τ — пороговый уровень отсечения модельной вероятности неисполнения контракта (в долях единицы);

FN_FEE_i — сумма комиссионного вознаграждения, полученная за выдачу i-гарантии, с результатом классификации FN (в денежных единицах);

TN_FEE_i — сумма комиссионного вознаграждения, полученная за выдачу i-гарантии, с результатом классификации TN (в денежных единицах);

FN_AMT_i — сумма, выданной i-гарантии, с результатом классификации FN (в денежных единицах);

C — сумма расходов на дополнительные данные, используемые в модельной оценке вероятности неисполнения контракта по заданному портфелю гарантий (в денежных единицах).

Далее необходимо найти такой уровень отсечения модельной оценки, при котором сумма прибыли будет максимальной.

$$P(\tau)\to\max \qquad (3)$$

Из формулы (2) также видно, что прибыль P зависит не только от результата классификации по уровню от-

сечения τ, но и от суммы расходов *C* на получение дополнительных данных для модели. Как правило, такие расходы имеют переменный характер, и зависят от количества запросов на получение дополнительных данных, необходимых для расчета по модели (например, плата за 1 запрос кредитной истории в одном из бюро кредитных историй)

Изменяя значения таких параметров, как уровень отсечения τ и сумму расходов *C* можно выделить два основных направления поиска максимальной прибыли по моделям с учетом дополнительных расходов и без них.

Стратегия № 1 выражается следующими формулами:

$$P_1 = \max\left(\left[\begin{array}{cccc} P^{v1}_{\tau 1} & P^{v1}_{\tau 2} & \dots & P^{v1}_{\tau 100} \end{array}\right]\right) \quad (4)$$

где P^{v1}— прибыль по модели без дополнительных расходов на всем заданном портфеле;

τ — перцентиль модельной оценки S^{v1} по модели без дополнительных расходов;

$$P_2 = \max\left(\left[\begin{array}{cccc} P^{v2}_{\varphi 1} & P^{v2}_{\varphi 2} & \dots & P^{v2}_{\varphi 100} \end{array}\right]\right) \quad (5)$$

где P^{v2} — прибыль по модели с дополнительными расходами на части заданного портфеля, где для *i*-гарантии выполняется условие $S_i^{v1} > S^{v1}(\tau^*)$;

τ* — уровень отсечения модельной оценки вероятности, соответствующий максимальному значению прибыли P_1;

φ — перцентиль модельной оценки S^{v2} по модели с дополнительными расходами.

$$P_{f1} = P_1 + P_2 \quad (6)$$

где P_{f1} — суммарная прибыль для всего заданного портфеля.

Стратегия № 2 выражается следующими формулами:

$$M_1 = \begin{bmatrix} \max([P^{v1}_{\beta 1\tau 1} & P^{v1}_{\beta 1\tau 2} & \dots & P^{v1}_{\beta 1\tau 100}]) \\ \max([P^{v1}_{\beta 2\tau 1} & P^{v1}_{\beta 2\tau 2} & \dots & P^{v1}_{\beta 2\tau 100}]) \\ \dots & \dots & & \dots \\ \max([P^{v1}_{\beta 100\tau 1} & P^{v1}_{\beta 100\tau 2} & \dots & P^{v1}_{\beta 100\tau 100}]) \end{bmatrix} \quad (7)$$

где β — перцентиль суммы гарантии на всем заданном портфеле;

P^{v1} — прибыль по модели без дополнительных расходов на заданном портфеле, где сумма *i*-гарантии <= сумма гарантии, соответствующей перцентилю β

$$M_2 = \begin{bmatrix} \max([P^{v2}_{\beta 1\varphi 1} & P^{v2}_{\beta 1\varphi 2} & \dots & P^{v2}_{\beta 1\varphi 100}]) \\ \max([P^{v2}_{\beta 2\varphi 1} & P^{v2}_{\beta 2\varphi 2} & \dots & P^{v2}_{\beta 2\varphi 100}]) \\ \dots & \dots & & \dots \\ \max([P^{v2}_{\beta 100\varphi 1} & P^{v2}_{\beta 100\varphi 2} & \dots & P^{v2}_{\beta 100\varphi 100}]) \end{bmatrix} \quad (8)$$

где P^{v2} — прибыль по модели с дополнительными расходами на заданном портфеле, где сумма *i*-гарантии > сумма гарантии, соответствующей перцентилю β

$$P_{f2} = \max(M_1 + M_2) \quad (9)$$

Далее задача заключается в выборе стратегии путем сравнения P_{f1} и P_{f2}.

Реализация метода на примере портфеля гарантий исполнения контракта 44-ФЗ

Рассмотрим изложенный выше метод оценивания прибыльности гарантий, а также стратегии нахождения максимальной прибыли на заданном портфеле гарантий исполнения контракта 44-ФЗ, на примере двух моделей машинного обучения, разработанных с помощью алгоритма градиентного бустинга, реализованного в библиотеке LightGBM. [7]

1) модель, разработанная на общедоступных бесплатных данных (далее «lgbm_1»);

2) модель, разработанная на комбинации общедоступных бесплатных данных и дополнительных данных, полученных из одного из бюро кредитных историй на коммерческой основе (далее «lgbm_2»).

Доля гарантий в обучающей выборке на момент разработки моделей, размеченных как событие «неисполнение контракта» не превышала 1,50 %, поэтому в качестве метрики оценки ранжирующей способности модели была выбрана AUC ROC (площадь «area» под кривой «receiver operating characteristic») [6].

Таблица 1.

Качество моделей по метрике AUC ROC

	Lgbm_1	Lgbm_2
обучающая	0,85	0,86
контрольная	0,84	0,85

Значение метрики AUC ROC модели «lgbm_2» выше на 1 п.п., чем у модели «lgbm_1». Учитывая тот факт, что данная метрика является интегральной оценкой ранжирующей способности модели и не учитывает значение уровня отсечения модельной оценки, отражающее результат итоговой классификации [1], необходимо выполнить дополнительный сравнительный анализ, опираясь на значение прибыльности обеих моделей на заданном портфеле банковских гарантий.

На Рисунке 1 видно, что максимальная прибыльность на заданном портфеле банковских гарантий составляет 0,93% и достигается использованием модели «lgbm_2_0_rub» (модель «lgbm_2», но без учета стоимости запроса дополнительных данных). Очевидно также, что при-

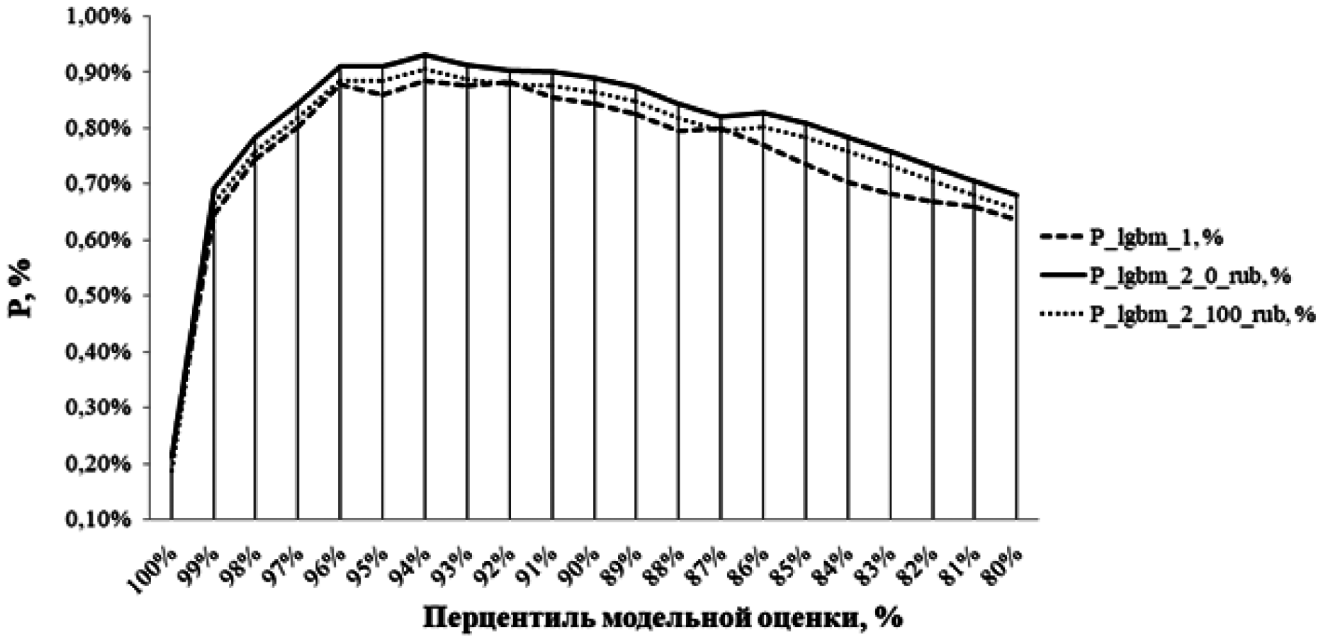

Рис. 1. Динамика прибыльности моделей на заданном портфеле в зависимости от перцентиля уровня отсечения модельной оценки

быльность модели «lgbm_2» чувствительна к изменению стоимости запроса дополнительных данных. Так, например, прибыльность модели «lgbm_2_100_rub» (модель «lgbm_2» с учетом стоимости запроса дополнительных данных 100 рублей за 1 запрос) соизмерима с прибыльностью, которую дает использование модели «lgbm_1».

Очевидно, что существует более эффективный способ управления расходами на запрос дополнительной информации для достижения максимальной прибыльности заданного портфеля при совместном использовании модели «lgbm_1» и «lgbm_2».

Выполним расчет прибыльности заданного портфеля при условии, что стоимость 1 запроса определяется значениями из массива [100, 200, 300, 400, 500, 700, 1000]. Результаты представлены в таблицах ниже.

Информация, представленная в Таблице 2, свидетельствует о том, что уровень модельной оценки модели «lgbm_1», соответствующий максимальной прибыльности, составляет 4,081 %, позволяет пропустить до выдачи 68,084 % от всего портфеля (с прибыльностью в 1,30 % на данном сегменте). Оставшаяся часть портфеля гарантий (31,916 %) отправлена на расчет модельной оценки по модели «lgbm_2».

Максимальная прибыльность, которую получилось достичь путем последовательного использования двух моделей на всем заданном портфеле по стратегии №1, составляет 0,910 % и достигается при стоимости 1 запроса, равной 100 рублей. При стоимости 1 запроса, составляющей 1000 рублей, прирост прибыльности от совместного использования двух моделей (в сравнении с использованием только модели «lgbm_1») составит

Таблица 2.

Расчет прибыльности заданного портфеля гарантий по стратегии №1

Цена 1 запроса, руб.	Уровень отсечения по lgbm_1, %	Прибыльность с учетом цены запросов lgbm_2, %	Доля портфеля lgbm_2, %	Прибыльность lgbm_1, %	Доля портфеля lgbm_1, %	Прибыльность всего, %
100	4,081 %	0,077 %	31,916 %	1,300 %	68,084 %	0,910 %
200	4,081 %	0,072 %	31,916 %	1,300 %	68,084 %	0,908 %
300	4,081 %	0,067 %	31,916 %	1,300 %	68,084 %	0,907 %
400	4,081 %	0,063 %	31,916 %	1,300 %	68,084 %	0,905 %
500	4,081 %	0,058 %	31,916 %	1,300 %	68,084 %	0,903 %
700	4,081 %	0,048 %	31,916 %	1,300 %	68,084 %	0,900 %
1000	4,081 %	0,034 %	31,916 %	1,300 %	68,084 %	0,896 %

Таблица 3.

Изменение прибыльности заданного портфеля гарантий от цены 1 запроса по стратегии №1

Уровень отсечения по lgbm_1, %	Цена 1 запроса, руб.	Прибыльность всего, %	Расходы на данные, %	Прирост прибыльности, %
—	—	0,885 %	—	—
4,081 %	100,00	0,910 %	0,002 %	0,025 %
4,081 %	200,00	0,908 %	0,003 %	0,023 %
4,081 %	300,00	0,907 %	0,005 %	0,022 %
4,081 %	400,00	0,905 %	0,006 %	0,020 %
4,081 %	500,00	0,903 %	0,008 %	0,018 %
4,081 %	700,00	0,900 %	0,011 %	0,015 %
4,081 %	1000,00	0,896 %	0,015 %	0,011 %

Таблица 4.

Расчет прибыльности заданного портфеля гарантий по стратегии № 2

Цена 1 запроса, руб.	Уровень отсечения по сумме гарантий, руб.	Прибыльность lgbm_2, %	Доля портфеля lgbm_2, %	Прибыльность lgbm_1, %	Доля портфеля lgbm_1, %	Прибыльность всего, %
100	335944	0,721 %	86,858 %	2,261 %	13,142 %	0,923 %
200	499882	0,668 %	82,660 %	2,115 %	17,340 %	0,919 %
300	499882	0,664 %	82,660 %	2,115 %	17,340 %	0,915 %
400	805035	0,620 %	76,200 %	1,846 %	23,800 %	0,912 %
500	4000000	0,567 %	41,465 %	1,153 %	58,535 %	0,910 %
700	4000000	0,564 %	41,465 %	1,153 %	58,535 %	0,909 %
1000	4000000	0,561 %	41,465 %	1,153 %	58,535 %	0,907 %

0,011 %, что меньше понесенных расходов на дополнительные данные для модели «lgbm_2».

При совместном использовании двух моделей по стратегии №2, максимальная прибыльность портфеля составит 0,923 % при стоимости 1 запроса, равной 100 рублей. При этом гарантии с суммой меньше или равно 335944 рубля направлены на модель «lgbm_1», а гарантии с суммой больше 335944 рублей — на модель «lgbm_2».

Информация, приведенная в Таблице 5, демонстрирует, что чувствительность прироста прибыльности по стратегии № 2 к величине расходов на дополнительные данные значительно ниже, чем по стратегии №1. При одинаковой стоимости в 100 рублей за 1 запрос, совместное использование двух моделей по стратегии №2, позволяет достичь прибыльность 0,923 %, что 0,013 % выше, чем по стратегии №1.

Заключение

Описанный в статье метод, а также предложенные стратегии совместного использования моделей машинного обучения позволяют найти максимальную прибыльность портфеля банковских гарантий, что в свою очередь может сделать процесс одобрения заявок на выпуск банковских гарантий более эффективным. Несмотря на то, что рассмотренные стратегии имеют разный результат, обе стратегии могут найти место в эффективной системе управления рисками, что положительно повлияет на финансовый результат банка, выпускающего гарантии.

Таблица 5.

Изменение прибыльности заданного портфеля гарантий от цены 1 запроса по стратегии №2

Уровень отсечения по сумме гарантии, руб.	Цена 1 запроса, руб.	Прибыльность всего, %	Расходы на данные, %	Прирост прибыльности, %
—	—	0,885%	—	—
335944	100,00	0,923%	0,005%	0,038%
499882	200,00	0,919%	0,007%	0,034%
499882	300,00	0,915%	0,011%	0,030%
805035	400,00	0,912%	0,010%	0,027%
4000000	500,00	0,910%	0,003%	0,025%
4000000	700,00	0,909%	0,004%	0,024%
4000000	1000,00	0,907%	0,005%	0,022%

ЛИТЕРАТУРА

1. Архипов В.А. Сравнительный анализ метрика качества для моделей бинарной классификации на примере кредитного скоринга // Вестник Алтайской Академии экономики и права. — 2019. — №9. — С.12–15.
2. Елисеев Д. Машинное обучение: прогнозирование рисков госзакупок / Елисеев Д., Романов Д. // Открытые системы. — 2018. — №2. — С.42–44.
3. Иванов Н.В. «Большие данные» и машинное обучение при управлении рисками невыполнения обязательств по контрактам в строительной отрасли / Иванов Н.В., Валпетерс М.Л., Киреев И.А. // Промышленное и гражданское строительство. — 2019. — №5. — С.81–87.
4. Кулаев М.А. Прогнозирование исполнения государственных контрактов на основе класса моделей машинного обучения — деревьев решений / Кулаев М.А, Домашова Д.В. // Материалы IV Международной научно-практической конференции международного сетевого института в сфере ПОД/ФТ. — 2019. — С.360–367.
5. Официальный сайт Единой информационной системы в сфере закупок. [Электронный ресурс]. — URL: https://zakupki.gov.ru/ (дата обращения: 20.10.2023)
6. Egan J. Signal detection theory and ROC analysis. — N.Y.: Academic press, 1975. — 277 p.
7. LightGBM. [Electronic resource]. — URL: https://lightgbm.readthedocs.io/en/latest/

Журнал «Современная наука: актуальные проблемы теории и практики»