АЛГОРИТМЫ БИННИНГА В МОДЕЛИРОВАНИИ КРЕДИТНОГО РИСКА

И. И. Севостьянова

Саратовский национальный исследовательский государственный университет им. Н. Г. Чернышевского, Россия
E-mail: irinasevostianova99@gmail.com

Статья посвящена особенностям использования алгоритмов биннинга для оптимизации моделирования кредитного риска. Актуальность исследования обусловлена существенным ростом кредитного риска для банковской сферы вследствие общей рецессии мировой экономики в период пандемии COVID-19. В качестве модели для оценки вероятности дефолта используется логистическая регрессия. Рассмотрены алгоритм разбиения на интервалы равной длины, алгоритм разбиения на равные по количеству наблюдений интервалы, монотонный биннинг. В рамках проведенного исследования предложена модификация указанных алгоритмов путем введения дополнительного этапа аналитической обработки. Работа выполнена на базе набора данных, описывающего годовую вероятность дефолта. Результатом исследования служит вывод о повышении качества оценки значения целевой переменной в случае проведения процедуры биннинга для информативных предикторов.

BINNING ALGORITHMS FOR CREDIT RISK MODELING

I. I. Sevostianova

The article is dedicated to the peculiarities of binning algorithms for credit risk modeling optimization. The relevance of the study is confirmed by a significant increase in credit risk for the banking sector due to the general recession in the global economy during the COVID-19 pandemic. Logistic regression is used as a base model for the probability of default estimation. Equal-size algorithm, equal-width algorithm and monotonous binning are considered. The modification of algorithms stated is proposed by introducing an additional stage of analytical processing. The study is performed on the dataset describing the annual probability of default (PD). The results of the work let us conclude that the quality of the target variable value estimation is improved in the case of applying the binning procedure for informative predictors.

Оценка кредитоспособности заемщиков является ключевой задачей в управлении кредитными рисками. Результаты анализа индивидуальных рисков составляют базу для оценки рисков всего кредитного портфеля и выстраивания эффективной стратегии работы финансового учреждения.

Задача процедуры биннинга состоит в проведении категоризации данных путем назначения каждой группе категории через ее вес (WOE – Weight Of Evidence). По итогам биннинга выполняется построение модели для оценки вероятности дефолта или определения значения целевой переменной. Поскольку зависимая переменная в задачах моделирования кредитного риска является чаще всего бинарной (0 – кредит выплачен, 1 – кредит не выплачен), построение модели в большинстве случаев выполняется на базе логистической регрес-

сии. Подобная закономерность обусловлена относительной легкостью интерпретации результатов моделирования, невысокой чувствительностью к выбросам, а также прямым моделированием вероятностей [1].

Алгоритмы биннига обеспечивают сегментацию фактора риска для статистически согласованных групп и повышают качество моделирования, что обуславливает непрерывное развитие методов категоризации. Так, например, в статье [2] предложен метод монотонного оптимального биннинга, а в работе [3] описана его модификация, которая для реального набора данных банковского учреждения обеспечила более высокую точность модели по сравнению с базовым алгоритмом, реализованным в библиотеке smbinning.

Качество биннинга в области моделирования кредитного риска оценивается по следующим критериям [4]:

1. отсутствующие значения группируются отдельно;
2. каждый бин должен содержать не менее 5% процентов наблюдений;
3. бины, содержащие только одинаковые значения для зависимой переменной, не могут быть использованы как результирующие.

Для количественной оценки эффективности алгоритмов биннинга используются показатели WOE и IV (Information Value). Веса категорий (WOE) позволяют найти границы чувствительности предиктора к появлению риска моделируемого события и провести оптимальным образом категоризацию количественных переменных. В свою очередь, показатель информативности (IV) отражает степень значимости предиктора для разделения значений целевой переменной [5]. Показатель IV используется в качестве основного показателя эффективности процедуры биннинга.

Расчетные формулы показателей приведены ниже:

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

где $d_i^{(1)}$, $d_i^{(2)}$ – относительные частоты «плохих» и «хороших» кредитов соответственно в i-том бине категоризированной переменной, $i = 1,\ldots,k$, k – число категорий переменной.

$$IV = \sum_{i=1}^{n}\left[WOE_i\left(d_i^{(1)} - d_i^{(2)}\right)\right]. \tag{2}$$

В проведенном исследовании можно выделить два ключевых этапа:

1. реализация, оптимизация и оценка базовых алгоритмов контролируемого и неконтролируемого биннинга;
2. построение и сравнительный анализ моделей прогнозирования дефолта с помощью метода логистической регрессии с учетом категоризации данных, проведенной по итогам процедуры биннинга.

Эмпирической основой исследования является набор данных, описывающий годовую вероятность дефолта, представленный в работе [6]. В качестве целевой переменной использовалась фиктивная бинарная переменная, которая принимает значение 1 (кредит был возвращен заемщиком, рейтинг заемщика высокий), в случае если у клиента отсутствует задолженность по кредит-

ному счету, клиент не является банкротом и кредитный счет клиента прошел первоначальный срок погашения без положительной остаточной задолженности. В противном случае переменной присваивается значение 0 (кредит не был возвращен, рейтинг заемщика низкий).

Среди рассмотренных предикторов следует отметить:

1. ежегодный доход (annual_income);

2 максимальное количество месяцев с просрочкой платежа за последние двенадцать месяцев (max_arrears_12m);

3. кредитный балл (bureau_score – рассчитывается по специальной модели исходя из данных о кредитной истории и текущем финансовом состоянии заемщика, характеризует вероятность возврата долга заемщиком);

4. величина текущей просрочки платежа в месяцах (arrears_months);

5. доля использования текущего кредитного счета (cc_util).

В рамках первого этапа исследования для оценки эффективности методов неконтролируемого биннинга были выбраны алгоритм разбиения на интервалы равной длины (equal-width) и алгоритм разбиения на равные по количеству наблюдений интервалы (equal-size). По результатам построения классических и модифицированных алгоритмов (модификация состояла в добавлении этапа аналитической обработки бинов, содержащих только одинаковые значения для целевой переменной), сделан вывод о низкой эффективности алгоритмов вследствие несоответствия базовым критериям качества биннинга.

В разрезе частных заключений, характерных для рассмотренного набора данных, следует отметить:

1. достижение более высокого уровня IV при использовании алгоритма разбиения на интервалы равной длины по сравнению с алгоритмом разбиения на равные по количеству наблюдений интервалы;

2. отсутствие монотонности WOE для алгоритма equal-size в преимущественной части проведенных экспериментов, свидетельствующее о некачественном биннинге;

3. высокая информативность предиктора, содержащего данные о максимальном количестве месяцев с просрочкой платежа за последний год.

В качестве исследуемого метода контролируемого биннинга был выбран алгоритм монотонного биннинга. Преимущество данного алгоритма заключается в более эффективной обработке несимметрично распределенных наборов данных. Для оценки качества биннига алгоритм был выполнен для 75% наблюдений исходного набора данных (обучающая часть), после чего по найденным точкам осуществлялось разбиение оставшихся 25% (тестовая часть). Показатели IV и WOE были рассчитаны как для тестовой части, так и для обучающей.

Оценка производительности алгоритма монотонного биннинга позволяет сделать вывод, что алгоритм является эмпирическим и требует дальнейшей аналитической обработки. Алгоритм демонстрирует относительно высокую эффективность по показателям IV, WOE, но не гарантирует удовлетворение главному условию качественного биннинга – наличию в каждом из результирующих бинов не менее 5% исходной выборки.

Второй и ключевой этап исследования состоял в построении и сравнительном анализе прогнозных моделей для определения значения зависимой переменной. Цель проведения экспериментов заключалась в оценке эффективности использования процедуры биннинга как одного из инструментов повышения качества модели.

Качество модели оценивалось на основе показателя accuracy и матрицы неточностей (confusion matrix). Сводная информация по результатам проведения экспериментов представлена в табл. 1, 2.

Таблица 1

Результаты построения прогнозных моделей

№ п/п	Алгоритм биннига	Регрессоры	Точность
1.	Логистическая регрессия *(биннинг не используется)*	cc_util, bureau_score, annual_income, max_arrears_12m, arrears_months	0.9589316
2.	Equal-size (по 4 бина)	cc_util (WOE), bureau_score (WOE), annual_income (WOE), max_arrears_12m (WOE), arrears_months (WOE)	0.9490505
3.	Equal-width (по 4 бина)	cc_util (WOE), bureau_score (WOE), annual_income (WOE), max_arrears_12m (WOE), arrears_months (WOE)	0.9563069
4.	Monotone	cc_util, bureau_score, annual_income**, max_arrears_12m (WOE),** arrears_months	**0.9606299**
5.	Monotone	cc_util, bureau_score, **annual_income (WOE),** max_arrears_12m, arrears_months	**0.9604755**
6.	Monotone	cc_util, **bureau_score (WOE),** annual_income, max_arrears_12m, arrears_months	0.9558438
7.	Monotone	cc_util (WOE), bureau_score (WOE), annual_income (WOE), max_arrears_12m (WOE), arrears_months (WOE)	0,9552
8.	Monotone	cc_util, bureau_score, annual_income, max_arrears_12m, **arrears_months (WOE)**	0.959858
9.	Monotone	cc_util, bureau_score, **annual_income (WOE)**, **max_arrears_12m (WOE)**, **arrears_months (WOE)**	**0.9603211**
10.	Monotone	cc_util, bureau_score, annual_income, **max_arrears_12m (WOE), arrears_months (WOE)**	**0.9618651**
11.	Monotone	cc_util, **bureau_score (WOE**), annual_income, **max_arrears_12m (WOE), arrears_months (WOE)**	**0.9635634**

Примечание. Приведены результаты только для лучших моделей среди серии экспериментов. Обозначение (WOE) указывает, для какого предиктора использовались не исходные значения, а найденные веса категорий.

Таблица 2

Значение лучшего показателя IV для предикторов (монотонный биннинг)

№ п/п	Предиктор	IV
1.	cc_util	4.63121
2.	bureau_score	0.41629
3.	annual_income	0.25358
4.	max_arrears_12m	1.061
5.	arrears_months	0.957

По результатам сравнительного анализа полученных моделей были сделаны следующие выводы:

1. использование в модели категоризированных предикторов, являющихся информативными с точки зрения значения показателя IV, обеспечивает повышение точности модели;

2. замена исходных значений на веса категорий для всех предикторов не позволяет обеспечить улучшение обобщающей способности модели, даже при условии значимости используемых независимых переменных, данная особенность объясняется наличием предикторов, для которых были выявлены явные признаки некачественного биннига (в данном случае cc_util);

3. алгоритмы контролируемого биннинга не продемонстрировали рост эффективности моделирования, однако разбиение на интервалы равной длины можно считать предпочтительным по сравнению с разбиением на равные по количеству наблюдений интервалы, как по результатам тестирования моделей, так и с точки зрения логики.

Автор благодарит доцента кафедры теории функций и стохастического анализа СГУ Агафонову Н. Ю. за постановку интересной задачи и внимание к работе.

СПИСОК ЛИТЕРАТУРЫ

1. *Kraus A.* Recent Methods from Statistics and Machine Learning for Credit Scoring / Munchen, 2014.

2. *Mironchuk P., Tchistiakov V.* Monotone optimal binning algorithm for credit risk modeling // [Электронный ресурс]. https://www.researchgate.net/publication/322520135. 2017. P. 1-15. (дата обращения: 01.10.2021).

3. *Агафонова Н. Ю., Козлов С. З.* Об одном методе оптимизации биннинга кредитных данных // Математическое моделирование и суперкомпьютерные технологии : сб. науч.тр. XX Межд. конф. под ред. проф. В. П. Гергеля. 2020. С. 27.

4. *Siddiqi N.* Credit risk scorecards: developing and implementing intelligent credit scoring / John Wiley and Sons, Inc., Hoboken, New Jersey, 2006.

5. *Сорокин А. С.* Построение скоринговых карт с использованием модели логистической регрессии // Науковедение. 2014. № 2. (21).

6. *Bellini T.* IFRS 9 and CECL Credit Risk Modelling and Validation: A Practical Guide with Examples // Worked in R and SAS. Academic Press. 2019. P. 654.