**УДК 004.89, 336**

# ОСОБЕННОСТИ ПОСТРОЕНИЯ СКОРИНГОВОЙ МОДЕЛИ НА ОСНОВЕ АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ DEDUCTOR

Лагерев Д.Г., Бондарева И.В.

Брянский государственный технический университет, Брянск, Россия

В статье рассматривается процесс построения скоринговой модели на основе данных АО «ОТП-Банк» с помощью аналитической платформы Deductor. Приводятся примеры необходимости использования скоринговых моделей в различных сферах деятельности. Описываются назначение и этапы разработки скоринговой модели. Рассматриваются особенности аналитической платформы Deductor и основные обработчики, необходимые для построения скоринговой модели. Подробно рассматриваются принципы действия обработчика «Конечные классы». Формируются выводы об особенностях построения модели на используемой платформе.

***Ключевые слова**: интеллектуальный анализ данных, скоринговая модель, Deductor, логистическая регрессия, классификация, конечные классы.*

**DOI:** 10.22281/2413-9920-2017-03-01-81-85

Идеи классификации популяции на группы в статистике были разработаны Фишером в 1936 г. на примере растений. В 1941 г. Дэвид Дюран впервые применил данную методику к классификации кредитов на «плохие» и «хорошие». Но широкое применение скоринга в области кредитования началось примерно в начале 60-х годов с распространением кредитных карточек.

Основная идея оценки риска банкротства распространилась посредством скоринг-моделей на другие аспекты кредитного риск-менеджмента:

- определение потенциальных (кредитоспособных) клиентов;
- определение приемлемых клиентов на стадии подачи заявки на кредит;
- определение возможного поведения текущих клиентов.

Скоринг начинают применять во многих сферах как совершенно новый подход к решению различных задач. В маркетинге скоринг используется для прогнозирования поведения участников рынка или клиентов. Построение скоринговых моделей позволяет получить ответы на такие вопросы, как: ответит ли потенциальный клиент на рекламное предложение, уйдет ли он к конкуренту, расторгнет ли договор, будут ли сбои в поставке товаров. Отдельное направление – использование скоринга в медицине в целях диагностики заболеваний по симптомам и результатам анализов.

Скоринг – это процедура классификации объектов в соответствии с их измеренными характеристиками, при этом неизвестен параметр, по которому разделяются группы, но известны другие факторы, связанные с интересующим параметром [1]. Скоринг используется во многих прикладных областях, в том числе и в банковской сфере. Скоринговая модель в общем смысле представляет собой статистическую модель, оценивающую вероятность наступления определенного события.

В упрощенном виде скоринговая модель представляет собой взвешенную сумму значений признаков, характеризующих потенциального заемщика. Интегральный показатель, получаемый на выходе модели, указывает степень риска, связанного с данным клиентом. Если учесть, что значение интегрального показателя, равное единице, указывает на максимальную надежность клиента, а нулевое значение – на некредитоспособность, то промежуточные значения позволяют сделать вывод о целесообразности выдачи кредита.

Скоринг – одна из прикладных областей интеллектуального анализа данных. Для получения эффективного решения можно использовать такие методы и модели, как нейронных сети, деревья решений, байесовская классификация, регрессионный анализ и другие.

Целью данной работы является построение скоринговой модели по данным банка АО «ОТП-Банк» [2]. В статье рассматривается один из этапов моделирования скоринговой модели, а именно определение конечных классов.

**Подход к решению задачи оценки рисков.** Проблема оценки кредитных рисков сводится к решению задач классификации и регрессии. К задаче классификации относится определение принадлежности объекта одной из заранее заданных категории риска, а к задаче регрессии – численная оценка вероятности возникновения неблагоприятного события. Для решения каждой из этих задач существует соответствующий математический аппарат, возможность применения которого зависит от данных, используемых для анализа.

Методика анализа структурированных данных о клиентах для оценки рисков состоит последовательности действий, состоящих в выдвижении гипотезы, сборе и систематизации данных в табличном виде, подборе модели, объясняющей имеющиеся прецеденты, интерпретации полученных результатов и применении полученной модели на новых данных для оценки риска. При этом следует учитывать, что необходима предварительная обработка данных с целью устранения пропусков и выбросов, объединения значений в интервалы и преобразования данных таким образом, чтобы в дальнейшем наиболее точно интерпретировать результат.

Подобный подход к анализу структурированных данных реализован в аналитической платформе Deductor [1]. Система обладает обширным функционалом и активно используется для создания решений в области анализа рисков. Платформа разработана таким образом, чтобы предоставить аналитику возможность сконцентрироваться на интеллектуальной работе и сделать процесс анализа данных полуавтоматическим.

**Основные этапы разработки скоринговой модели.** Необходима предварительная обработка данных, которая заключается в получении данных, устранении выбросов и пропущенных значений. За предварительной обработкой следует моделирование скоринговой карты.

Выделим основные этапы моделирования:

- сэмплинг;
- двумерный анализ (или: категоризация, оптимальное квантование);
- расчет весов и баллов при необходимости;
- оценка качества модели.

В качестве источника данных для анализа выступает скоринговая анкета банка АО «ОТП-Банк» [2], который входит в число 50 крупнейших банков России. Цель анализа – предсказание отклика клиентов. Исходная выборка содержит записи о 15 223 клиентов, классифицированных на два класса: 1 – отклик был (1812 клиентов), 0 – отклика не было (13411 клиентов). Ещё 14910 записей отложены в качестве тестовых. Записи (признаковые описания) клиентов состоят из 50 признаков, в состав которых входит, в частности, возраст, пол, социальный статус относительно работы, социальный статус относительно пенсии, количество детей, количество иждивенцев, образование, семейное положение, отрасль работы.

**Обработчик Конечные классы.** Рассмотрим этап построения скоринговой модели, на котором необходимо уменьшить число значений исходного набора данных. Формирование конечных классов производится с целью предобработки выборок для повышения качества логистической регрессии [3]. С помощью обработчика Конечные классы [1] можно выполнить данную задачу за счет объединения значений в пределах некоторого интервала с использованием информации о бинарной выходной переменной.

Данный обработчик предназначен для решения следующих задач:

- снижение разнообразия значения признаков без ущерба для информативности данных;
- снижение размерности данных за счет исключения признаков с низкой значимостью;
- восстановление пропусков;
- борьба с выбросами и экстремальными значениями;
- упрощение описания исследуемых объектов.

Две цели оптимального квантования:

- стремление к простоте;
- стремление минимизировать потерю информации.

Ограничения для конечных классов:

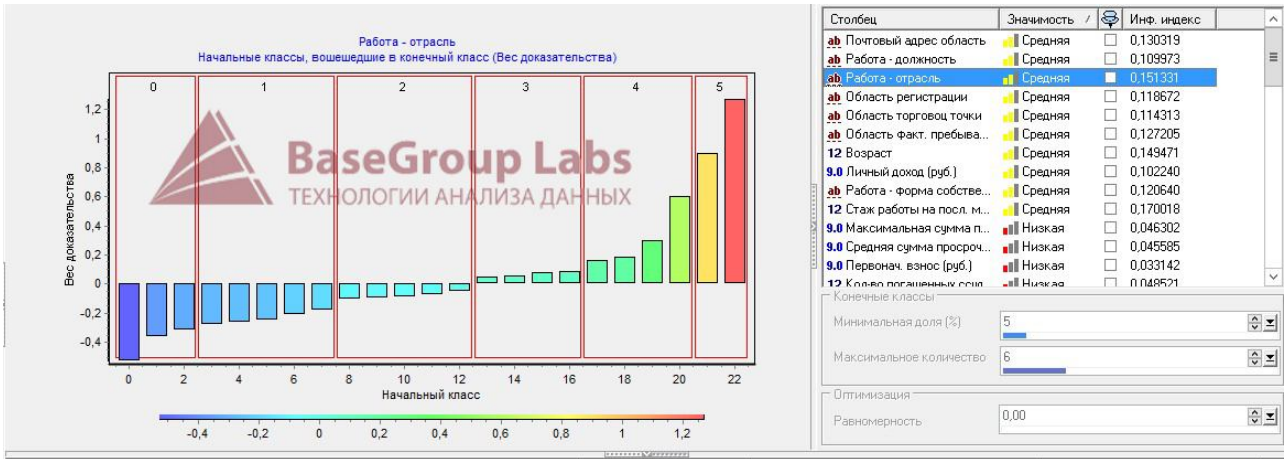
- минимальное число наблюдений в конечном классе;
- максимальное число конечных классов.

Процедура сокращения уникальных значений признака осуществляется следующим образом:

1. Формируется исходное множество уникальных значений поля до обработки или *начальные классы* (fine classing);
2. Происходит «сжатие» начальных классов в меньшее количество интервалов, называемых *конечными классами* (coarse classing).

В обработчике заданы значения по умолчанию параметров Минимальная доля и Максимальное количество, равные 5% и 6 соответственно. Пользователь может задать другие значения до выполнения расчетов данного обработчика либо в интерактивном режиме.

Использование интерактивного режима позволяет редактировать количество конечных классов и их границы. После каждого изменения происходит автоматический перерасчет целевой функции, веса доказательства и информационного индекса (рис. 1). Рассмотрим процесс вычисления информационного индекса.




| Конечные классы | | Границы | | Целевая функция | | | | Критерий | |
|---|---|---|---|---|---|---|---|---|---|
| Номер | Метка | Нижняя | Верхняя | События | Не-события | Всего | Доля | Вес доказательства | Инф. индекс |
| 0 | [Ресторанный бизнес/Общественное питание; Торговля; Транспорт] | Минимум | Транспорт... | 586 | 2994 | 3580 | 0,235 | -0,370598 | 0,037115 |
| 1 | [Сборочные производства; Развлечения/Искусство; ЧОП/Детективная д-ть; Коммуна... | Транспорт... | Информац... | 158 | 932 | 1090 | 0,072 | -0,226906 | 0,004017 |
| 2 | [Банк/Финансы; Информационные технологии; Другие сферы; Химия/Парфюмерия/... | Информац... | Нефтегазо... | 313 | 2133 | 2446 | 0,161 | -0,082563 | 0,001130 |
| 3 | [Сельское хозяйство; Государственная служба; Образование; Металлургия/Промышл... | Нефтегазо... | Металлург... | 486 | 3856 | 4342 | 0,285 | 0,069533 | 0,001343 |
| 4 | [Здравоохранение; Строительство; Салоны красоты и здоровья; Энергетика] | Металлург... | Энергетик... | 198 | 1797 | 1995 | 0,131 | 0,203963 | 0,005043 |
| 5 | [Наука; Не задано] | Энергетик... | Максимум | 71 | 1699 | 1770 | 0,116 | 1,173471 | 0,102683 |
| Всего | | | | 1812 | 13411 | 15223 | | | 0,151331 |

Рис. 1. Визуализатор Конечные классы

Для формирования конечных классов используется $Wo$-анализ (weight of evidence) или совокупность доказательства – статистический метод оценки влияния тех или иных факторов на справедливость некоторой гипотезы. Используется гипотеза независимого поведения признаков, которая заключается в следующем: пропорция события и не-события в анализируемой подгруппе должна сохраняться такой же, как и для всей выборки в целом.

Диапазон изменений признака разбивается на несколько начальных классов, для каждого из которых вычисляется коэффициент по формуле:

$$WoE_i = \ln \frac{N_i / N}{P_i / P},$$

где $i$ – индекс признака, для которого вычисляется показатель $WoE$; $N_i$ – число не-событий в $i$-й группе; $N$ – общее число не-событий; $P_i$ – число событий в $i$-й группе; $P$ – общее число событий.

Если значение категории совпадает с событием большее число раз, чем с не-событием, то согласно формуле, под знаком логарифма будет значение меньше 1, что делает его отрицательным. Значение $WoE < 0$ указывает на большую вероятность появления собы-

тия, а $WoE > 0$ – не-события. Индекс $WoE$ есть количественная мера предиктивной силы отдельной категории внутри переменной.

$WoE$ является промежуточным элементом для вычисления агрегированной величины, называемой информационным индексом $IV$ (Information Value):

$$IV = \sum_{i=1}^{K} \left\{ \left( \frac{N_i}{N} - \frac{P_i}{P} \right) \cdot WoE_i \right\}.$$

Информационный индекс всегда является положительной величиной и отвечает за предсказательную способность всей переменной.

Коэффициенты $WoE$ и вычисленные на их основе значения $IV$ являются критерием для формирования конечных классов оптимальным образом:

- максимизация значимость признака в бинарной классификационной модели;
- максимизация равномерность заполнения интервалов, что обеспечивает наилучшую репрезентативность результатов;
- сочетание данных вариантов.

**Логистическая регрессия.** В целях обеспечения адаптивности модели оценки риска следует использовать алгоритм, предусматривающий возможность подстройки под вносимые изменения. Рассмотрим логистическую регрессию, назначение которой состоит в анализе связи между несколькими независимыми переменными и зависимой переменной. Можно рассматривать линейную модель для задач с бинарным результирующим полем (выдать кредит либо отказать в выдаче кредита). На основе такого алгоритма строятся скоринговые карты, позволяющие подобрать оптимальный и экономически обоснованный порог отсечения.

Логистическая кривая имеет вид

$$P = \frac{1}{1 - e^{-y}},$$

где $p$ – вероятность того, что произойдет интересующее событие; $y$ – стандартное уравнение регрессии:

$$y = \ln\left( \frac{P}{1-P} \right) - \beta_1 + \beta_1 x.$$

Преобразование $\ln\left( \frac{P}{1-P} \right)$ называется логистическим или логитом. Применение логит-преобразования позволяет предсказывать непрерывную переменную со значениями на отрезке [0; 1].

Вычисление требуемых коэффициентов, информационного индекса и значений целевой функции без использования специального программного обеспечения представляет собой трудоемкий процесс. Поскольку платформа Deductor позволяет осуществлять быстрый перерасчет всех необходимых значений, возможно выполнять анализ данных итерационно.

После построения очередной скоринговой модели следует сравнить эффективность моделей. Результаты проверки на тестовых данных могут не совпадать с действительным результатом, поэтому основной целью является проверка на реальных данных

В результате анализа функциональных возможностей платформы Deductor и применении их в построении скоринговой модели можно сделать вывод, что грамотный выбор множества значений для входных переменных повышает качество и скорость построения модели. Отбор переменных для модели в ходе оценки индекса $WoE$ обеспечивает наибольшую репрезентативность результатов, делает скоринговую модель устойчивой и также повышает скорость ее построения. Все этапы моделирования можно выполнить без использования специальных программных средств, но использование Deductor позволяет аналитику сконцентрироваться на интеллектуальной работе и сделать процесс анализа данных полуавтоматическим. Таким образом, возможно сократить время на создание и повысить качество скоринговой модели.

**Список литературы**

1. BaseGroup Labs ООО «Аналитические технологии» [Электронный ресурс]. – Режим доступа: https://basegroup.ru/deductor, свободный.
2. АО «ОТП-Банк» [Электронный ресурс]. – Режим доступа: https://www.otpbank.ru/.
3. Паклин, Н.Б. Оптимальное квантование для повышения качества бинарных классификаторов / Н.Б. Паклин, В.В. Афанасьев // Искусственный интеллект. – 2013. – Вып. 4. – С. 392-399.

**Сведения об авторах**

Лагерев Дмитрий Григорьевич – кандидат технических наук, доцент кафедры «Информатика и программное обеспечение» ФГБОУ ВО «Брянский государственный технический университет», *LagerevDG@yandex.ru*.

Бондарева Инна Васильевна - магистрант по направлению «Информатика и вычислительная техника» ФГБОУ ВО «Брянский государственный технический университет», *innagorda@ya.ru*.

## CONSTRUCTION FEATURES OF SCORING MODELS BASED ON DEDUCTOR ANALYTICAL PLATFORM

Lagerev D.G., Bondareva I.V.

Bryansk State Technical University, Bryansk, Russian Federation

The aim of the study is to build a scoring model based on Deductor analytical platform. Problems solved by this method are introduced to actualizing the research. The source data for customer response prediction is the scoring profile of the OTP Bank. Modeling process and Fine&Coarse Classing Deductor's handler are described. Fine&Coarse Classing is capable of reducing the unique values and allows to analyze iteratively and to improve the quality and speed of model building. The weight of evidence, information value and the value of the objective function enable the fundamental parameters to be determined from experimental results. The conclusion is made that using Deductor have several advantages over manual calculations.

***Keywords**: Data mining, scoring model, deductor, classification, logistic regression, Fine&Coarse Classing.*
**DOI:** 10.22281/2413-9920-2017-03-01-81-85

**References**

1. BaseGroup Labs OOO Analytical technology. Available at: https://basegroup.ru/deductor.
2. AO OTP-Bank. Available at: https://www.otpbank.ru/.
3. Paklin N.B., Afanasiev V.V. Optimal Quantization to Improve the Quality of Binary Classifiers. *Artificial intelligence*, 2013, No.4, pp. 392-399.

**Authors' information**

Dmitriy G. Lagerev - Candidate of Technical Sciences, Associate professor of the Departament "Computer Science and Engineering" at Bryansk State Technical University, *LagerevDG@yandex.ru*.

Inna V. Bondareva - undergraduate of the faculty "Computer Science and Engineering" at Bryansk State Technical University, *innagorda@ya.ru.*

Дата публикации
(Date of publication):
25.03.2017