НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

ИНСТИТУТ ПРОБЛЕМ УПРАВЛЕНИЯ
им. В.А. ТРАПЕЗНИКОВА РАН

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2021)

МАТЕРИАЛЫ
XX Международной конференции
имени А. Ф. Терпугова
1–5 декабря 2021 г.

ТОМСК
Издательство Томского
государственного университета
2022

УДК 519
ББК 22.17
И74
Информационные технологии и математическое моделирование (ИТММ-2021): Материалы XX Международной конференции имени А. Ф. Терпугова (1–5 декабря 2021 г.). — Томск: Издательство Томского государственного университета, 2022. — 391 с.

ISBN 978–5–907572–20–1

Сборник содержит избранные материалы XX Международной конференции имени А.Ф. Терпугова по следующим направлениям: теория массового обслуживания и ее приложения, интеллектуальный анализ данных и визуализация, информационные технологии и программная инженерия, математическое и компьютерное моделирование технологических процессов.

Для специалистов в области информационных технологий и математического моделирования.

УДК 519
ББК 22.17

Конференция проведена при поддержке
международного научно-методического центра
Томского государственного университета по математике,
информатике и цифровым технологиям в рамках
федерального проекта «Кадры для цифровой экономики»
национальной программы
«Цифровая экономика в Российской Федерации»

ISBN 978–5–907572–20–1

NATIONAL RESEARCH TOMSK STATE UNIVERSITY

PEOPLES' FRIENDSHIP UNIVERSITY OF RUSSIA

V.A. TRAPEZNIKOV INSTITUTE OF CONTROL
SCIENCES OF RUSSIAN ACADEMY OF SCIENCES

INFORMATIONAL TECHNOLOGIES AND MATHEMATICAL MODELLING (ITMM-2021)

PROCEEDINGS
of the 20th International Conference
named after A. F. Terpugov
2021 December, 1–5

TOMSK
Tomsk State
University Publishing
2022

UDC 519
LBC 22.17
I60
Informational technologies and mathematical modelling (ITMM-2021): Proceedings of the 20th International Conference named after A. F. Terpugov (2021 December, 1–5). — Tomsk: Tomsk State University Publishing, 2021. — 391 p.

ISBN 978–5–907572–20–1

This volume presents selected papers from the XIX International Conference named after A.F. Terpugov. The papers are devoted to new results in the following areas: queuing theory and its applications, data mining and visualization, information technology and software engineering, mathematical and computer modeling of technological processes.

UDC 519
LBC 22.17

E d i t o r s:
A.A. Nazarov, Doctor of Technical Sciences, Professor,
S.P. Moiseeva, Doctor of Physical and Mathematical Sciences, Professor,
A.N. Moiseev, Doctor of Physical and Mathematical Sciences, Associate Professor.

The conference was supported by
International Computer Science
Continues Professional Development Center
of the Federal project "Human Resources for the Digital Economy"
of the National program
"Digital Economy of the Russian Federation"

ISBN 978–5–907572–20–1

Методы анализа и визуализации данных

ОСНОВНЫЕ ЭТАПЫ ОБРАБОТКИ И МЕТОДЫ ОТБОРА ПРИЗНАКОВ ДЛЯ ДАЛЬНЕЙШЕГО ПРОГНОЗИРОВАНИЯ ПЛАТЕЖЕСПОСОБНОСТИ КЛИЕНТОВ БАНКА

Д. Д. Бугакова[1], Е. Ю. Лисовская[1], Г. В. Баймеева[2]

[1] *Национальный исследовательский Томский государственный университет, г. Томск, Россия,*
[2] *X5 Retail Group, г. Москва, Россия*

В данной работе рассматриваются основные этапы обработки и методы отбора признаков для их дальнейшего использования в алгоритмах машинного обучения для построения моделей, которые предназначены для прогнозирования платежеспособности клиентов банка. В работе были рассмотрены такие способы отбора признаков как: расчет коэффициентов WoE (Weight of Evidence) для признаков с последующей оценкой предсказательной силы отобранных факторов с помощью расчета коэффициента IV (information value) и оценка важности признаков с помощью алгоритма случайного леса совместно с методом RFE (recursive feature elimination), основанного на логистической регрессии.

Ключевые слова: *Обработка данных, WoE (Weight of Evidence), IV (information value), квантование (биннинг), логистическая регрессия.*

Введение

Перед обучением модели на данных предварительно их нужно обработать. Обычно, обработка данных состоит из двух этапов: первичная обработка данных (удаление выбросов, удаление дубликатов из выборки, замена или удаление пропущенных значений), отбор информативных признаков (так как для обучения моделей необходимы не все признаки, а только те, которые в большей степени влияют на целевую признак).

1. Обзор исследуемых данных

Для того, чтобы на практике посмотреть работу алгоритмов воспользуемся набором данных с сайта kaggle.com о клиентах банка «Тинькофф» [1], для которого предлагается по данным из анкеты с исполь-

зованием алгоритмов машинного обучения предсказать факт наличия дефолта.

Набор данных содержит информацию о 205296 клиентах и 17 признаках, 5 количественных (возраст, скоринговый балл, количество обращений в банк, доход, количество отказанных заявок), 7 категориальных (количество связей с другими клиентами, уровень образования, пол, количество лет, которое заявитель является клиентом банка, тип жилплощади, должность, регион), 4 бинарных признаках (наличие автомобиля, иномарки, дохода выше среднего, загранпаспорта) и 1 целевого признака (наличие дефолта у клиента).

2. Первичный анализ данных

После проведения первичного анализа данных был удален признак «даты подачи заявления», т.к. он не несет в себе никакой полезной информации для дальнейшей работы, но возможно этот признак можно было исследовать на сезонность и не удалять этот признак сразу, удалены дубликаты. Были обнаружены и обработаны пропуски во всем наборе данных. Для признака «количество отказанных заявок» была проведена проверка на информативность. Количественный признак будем считать неинформативным, если в нем большинство строк с одинаковыми значениями. Выбранный признак оказался на 82% неинформативным следовательно его можно удалить. В некоторых признаках были обнаружены выбросы на основе межквартильного размаха, значения, которые не попали в этот отрезок были удалены.

3. Отбор признаков

Для лучшего определения того, какие признаки влияют на вероятность дефолта клиента, нужно рассмотреть взаимоотношения между целевым признаком и остальными признаками. На основе графического представления зависимостей были выделены признаки, которые оказывают большее влияние на значение целевого признака (Рис. 1).

После первичного анализа данных, нужно осуществить отбор признаков, потому что для обучения алгоритмов нужны не все признаки, а только те, которые в большей степени влияют на итоговый результат.

Для отбора признаков будут использованы следующие методы:

1) WoE (Weight of Evidence) с последующей оценкой предсказательной силы отобранных факторов с помощью алгоритма IV (information value) [2];
2) Оценка важности признаков с помощью алгоритма случайного леса.

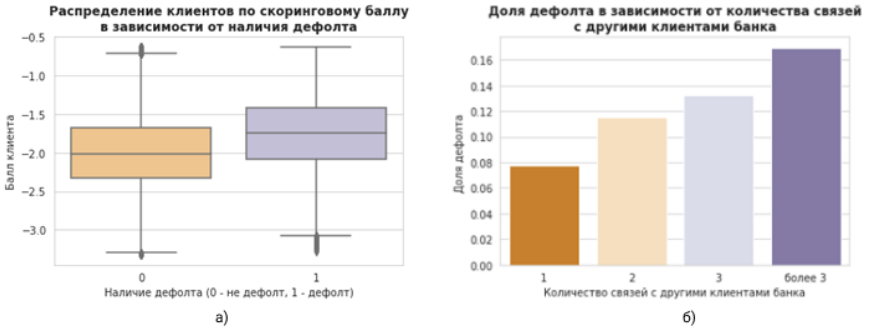

Рис. 1. а) График распределения клиентов по скоринговому баллу в зависимости от наличия дефолта, б) Доля дефолта в зависимости от количества связей с другими клиентами банка

Перед тем как начать работу по отбору признаков первым способом нужно все количественные непрерывные признаки преобразовать в категориальные, используя квантование (биннинг) [4]. Квантование – это процесс обработки данных, который позволяет разбить диапазон количественного признака на заданное количество интервалов (бинов) и присвоить каждому бину название. Квантование бывает двух видов:

1) *Интервальное*. Диапазон значений делится на одинаковые интервалы, в каждом не будет слишком много или мало данных.
2) *Квантильное*. Ширина интервалов будет различна, но в каждый попадет примерно одинаковое количество значений.

В настоящей работе применяется второй вид квантования. С помощью квантования будут преобразованы следующие признаки: возраст клиента, скоринговый балл и доход клиента.

4. Отбор признаков первым способом

Коэффициент WoE относительно данной задачи, характеризует степень отклонения уровня дефолтов в данной группе от среднего значения в выборке.

Для расчета WoE нужно для каждого категориального признака и для каждой группы внутри признака вычислить число клиентов с дефолтом («плохие» клиенты) и без дефолта («хорошие» клиенты) и рассчитать коэффициент по следующей формуле:

$$WoE_i = \ln\left(\frac{p_i}{q_i}\right),$$

где i – номер группы внутри признака, p – доля «хороших» клиентов среди всех «хороших», q – доля «плохих» клиентов среди всех «плохих».

После расчета WoE рассчитывается информационная ценность (коэффициент IV), который характеризует статистическую значимость признака, по следующей формуле:

$$IV = \sum_{i=1}^{n} (p_i - q_i) * WoE_i.$$

Для определения предсказательной силы признака на основе расчета IV воспользуемся следующей классификацией (Значение IV – Предсказательная сила):

— $< 0,02$ – отсутствует,
— $0,02 - 0,1$ – низкая,
— $0,1 - 0,3$ – средняя,
— $> 0,3$ – высокая.

Для признаков, которые были выбраны как наиболее значимые ниже представлена таблица 1 со значениями коэффициента IV. После отбора признаков созданы фиктивные признаки с помощью dummy-кодирования [3].

Пусть один из признаков x_j принимает m значений $\{b_1, \dots b_m\}$, тогда для каждого объекта x^j можно заменить признак x_i^j на $m-1$ признаков со значениями $\{0, 1\}$:

$$Z_i^{b_k} = I\left[x_i^j = b_k\right],\ k \in \{1, \dots, m-1\},$$

где $I[A]$ – индикатор события A.

Таблица 1

Значения коэффициента IV для наиболее значимых признаков

Признак	Значение IV
Количество связей с другими клиентами	0,13
Скорринговый балл	0,27
Тип жилплощади	0,11
Тип занимаемой должности	0,10
Число лет, которое заявитель является клиентом	0,10

Последним шагом в отборе признаков первым способом является удаление сильно коррелирующих между собой признаков (из двух при-

знаков нужно оставить те, у которых значение коэффициента WoE больше), для этого были построены матрицы корреляций (Рисунок 2).

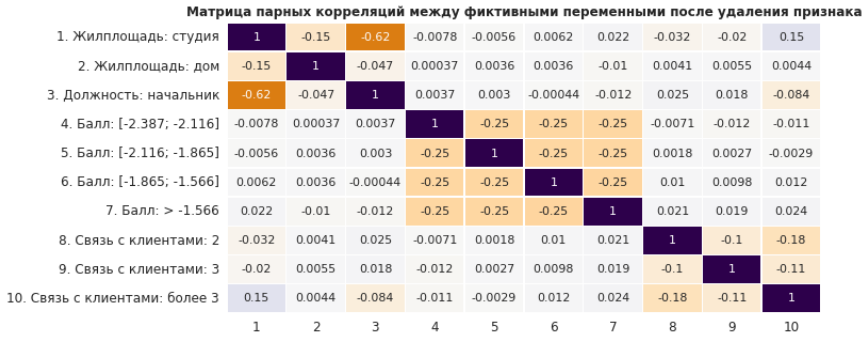

Рис. 2. Матрица корреляций между фиктивными признаками после удаление сильно коррелирующих признаков

Таким образом итоговый набор данных состоит из 10 отобранных первым способом признаков: Жилплощадь: студия; Жилплощадь: дом; Должность: начальник; Балл: [-2.387; -2.116]; Балл: [-2.116; -1.865]; Балл: [-1.865; -1.566]; Балл: больше -1.566; Связь с клиентами: 2; Связь с клиентами: 3; Связь с клиентами: более 3.

5. Отбор признаков вторым способом

Второй способ оценивает важность каждого признака на основе алгоритма случайного леса. Для этого нужно обучить модель на тренировочной выборке и посчитать out-of-bag ошибку для каждого объекта этой выборки. Ошибка усредняется для каждого элемента по всему случайному лесу. Значения каждого признака перемешиваются для всех объектов обучающей выборки и вычисления ошибки производятся заново, чтобы оценить важность признака. Чем больше уменьшается точность предсказаний из-за исключения или перестановки признака, тем важнее этот признак.

$$FI^{(t)}(x_j) = \frac{\sum_{i \in OOB^{(t)}} I\left(y_i = \widehat{y}_i^{(t)}\right)}{|OOB^{(t)}|} - \frac{\sum_{i \in OOB^{(t)}} I\left(y_i = \widehat{y}_{i,\pi_j}^{(t)}\right)}{|OOB^{(t)}|},$$

где $OOB^{(t)}$ – out-of-bag ошибка для дерева $t \in \{1, \ldots, N\}$, N – количество деревьев в случайном лесе, x_j – признак, для которого оценивается

важность, $\widehat{y}_i^{(t)}$ – предсказание перед удалением или перестановкой признака, $\widehat{y}_{i,\pi_j}^{(t)}$ – предсказание после удаления или перестановки признака.

Далее производится расчет важности признака по всем деревьям в случайном лесе и может быть представлен в двух формах: ненормализованной и нормализованной

$$FI(x_j) = \frac{1}{N}\sum_{t=1}^{N} FI^{(t)}(x_j),\ z_j = \frac{N \cdot FI(x_j)}{\sigma},$$

где σ – стандартное отклонение разностей.

В таблице 2 приведены расчёты коэффициентов для пяти наиболее важных признаков. Для дальнейшего анализа было выбрано 24 признака для которых нормализованное значение больше 0,02.

На выбранных признаках был использован метод рекурсивного сокращения RFE (recursive feature elimination) в сочетании с логистической регрессией. В данной работе этот метод был использован как дополняющий, но также может быть использован, как и самостоятельный.

Таблица 2

Важность признаков

Признак	Ненормализованное значение	Нормализованное значение
Регион: МСК	227,8	0,043968
Жилплощадь: студия	207,4	0,040031
Пол: мужской	199,0	0,038410
Балл: >-1.566	189,4	0,036557
Балл: [-1.865; -1.566]	182,1	0,035148

Суть метода: модель обучается на исходном наборе признаков, оценивает их значимость и исключает наименее важный признак, процесс повторяется до тех пора пока не будет получено оптимальное или заданное количество признаков, каждому признаку присваивается ранг, чем выше ранг, тем важнее признак.

Таким образом итоговый набор данных состоит из 10 отобранных вторым способом признаков: Образование: ВШ; Жилплощадь: студия; Балл: [-2.116; -1.865]; Балл: [-1.865; -1.566]; Балл: > -1.566; Связь с клиентами: более 3; Наличие автомобиля; Клиент банка: более 3; Регион: СПБ; Регион: МСК.

Заключение

После рассмотрения методов по отбору признаков было установлено, что оба метода отбирают примерно одинаковые признаки. Для оценки работы методов в будущем предлагается построить модели на данных отобранных двумя методами, сравнить их, используя метрики качества и сделать вывод о работе моделей на данных, отобранных разными методами. Также предлагается проверить качество моделей на отобранных данных после применения метода undersampling, который предназначен для балансировки данных.

СПИСОК ЛИТЕРАТУРЫ

1. `https://www.kaggle.com/c/fintech-credit-scoring.` — Kaggle. 2021.
2. *Шунина Ю. С.* Критерии качества работы классификаторов. // Вестник Ульяновского государственного технического университета. 2015. № 2. С. 67–70.
3. *Карминский А. М.* Сравнительный анализ методов прогнозирования банкротств российских строительных компаний. // Бизнес-информатика. 2019. № 3. С. 52–66.
4. *Шунина Ю. С.* Прогнозирование кредитоспособности клиентов на основе методов машинного обучения. // Финансы и кредит. 2015. № 27. С. 2–12.

Бугакова Дарья Дмитриевна — студент института прикладной математики и компьютерных наук. E-mail: *bugashka17@inbox.ru*

Лисовская Екатерина Юрьевна — к.ф.-м.н., доцент кафедры теории вероятностей и математической статистики института прикладной математики и компьютерных наук. E-mail: *ekaterina_lisovs@mail.ru*

Баймеева Галина Владимировна — старший менеджер по работе с большими данными. E-mail: *baymeevag@gmail.com*

СОДЕРЖАНИЕ

Научное издание

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ
(ИТММ-2021)

МАТЕРИАЛЫ
XX Международной конференции
имени А. Ф. Терпугова
1–5 декабря 2021 г.

Редактор *В.Г. Лихачева*
Компьютерная верстка *Д.В. Семенова, Е.Ю. Лисовская, О.Д. Лизюра*
Дизайн обложки *Л.Д. Кривцовой*

Отпечатано на оборудовании
Издательства Томского государственного университета
634050, г. Томск, пр. Ленина, 36.
Тел. 8+(382-2)–52-98-49
Сайт: http://publish.tsu.ru
E-mail: rio.tsu@mail.ru

Подписано к печати 12.09.2022 г.
Формат $60 \times 84^1/_{16}$. Бумага для офисной техники. Гарнитура «Times».
Печ. л. 24.5. Усл. печ. л. 22.7. Тираж 500 экз. Заказ № 5145.

ISBN 978-5-907572-20-1

ИНФОРМАЦИЯ О ПУБЛИКАЦИИ

eLIBRARY ID: 49546401 EDN: TFAVMA

ОСНОВНЫЕ ЭТАПЫ ОБРАБОТКИ И МЕТОДЫ ОТБОРА ПРИЗНАКОВ ДЛЯ ДАЛЬНЕЙШЕГО ПРОГНОЗИРОВАНИЯ ПЛАТЕЖЕСПОСОБНОСТИ КЛИЕНТОВ БАНКА

БУГАКОВА Д. Д.[1], ЛИСОВСКАЯ Е. Ю.[1], БАЙМЕЕВА Г. В.[2]

[1] Национальный исследовательский Томский государственный университет
[2] X5 Retail Group

Тип: статья в сборнике трудов конференции Язык: русский Год издания: 2022

Страницы: 56-62

ИСТОЧНИК:

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2021)
материалы XX Международной конференции имени А. Ф. Терпугова. Томск, 2022
Издательство: Национальный исследовательский Томский государственный университет

КОНФЕРЕНЦИЯ:

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2021)
Томск, 01–05 декабря 2021 года
Организаторы: Национальный исследовательский Томский государственный университет

КЛЮЧЕВЫЕ СЛОВА:

ОБРАБОТКА ДАННЫХ, КВАНТОВАНИЕ (БИННИНГ), ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

АННОТАЦИЯ:

В данной работе рассматриваются основные этапы обработки и методы отбора признаков для их дальнейшего использования в алгоритмах машинного обучения для построения моделей, которые предназначены для прогнозирования платежеспособности клиентов банка. В работе были рассмотрены такие способы отбора признаков как: расчет коэффициентов WoE (Weight of Evidence) для признаков с последующей оценкой предсказательной силы отобранных факторов с помощью расчета коэффициента IV (information value) и оценка важности признаков с помощью алгоритма случайного леса совместно с методом RFE (recursive feature elimination), основанного на логистической регрессии.

БИБЛИОМЕТРИЧЕСКИЕ ПОКАЗАТЕЛИ:

Входит в РИНЦ®: да
Входит в ядро РИНЦ®: нет
Цитирований в РИНЦ®: 0
Цитирований из ядра РИНЦ®: 0