14. Семенова, Е.Г. Взвешивание иерархии показателей оценки качества программно-аппаратных комплексов данных / С.А. Морозов, Я.А. Ивакин, Е.Г. Семенова, М.С. Смирнова // Научный журнал «Вестник Санкт-Петербургского государственного университета технологии и дизайна» №5, 2017. С. 136-143
15. Смирнова, М.С. Обеспечение качества программно-аппаратных комплексов для центров хранения и обработки данных / С.А. Морозов, В.М. Балашов, М.С. Смирнова // Вопросы радиоэлектроники. 2018. №3 С. 145-150
16. Дюваль П.М. Непрерывная интеграция. Улучшение качества программного обеспечения и снижение риска [Текст] Дюваль П.М., Матиас С., Гловер Э. – СПб.: Символ, 2016.- 240с.
17. Watson, D.F. A Guide to the Analysis and Display of Spatial Data [Text] / D.F. Watson. - Oxford Pergamum Press, 2009. – 321 p.
18. Walford, N. Geographical Data Analysis / N. Walford. – N. Y.: John Wiley & Sons. - 2015.
19. White, F.E. A Model for Data Fusion / F.E. White // 1st National Symposium on Sensor Fusion: Proc. – 2012.

МЕТОДОЛОГИЯ ПОДГОТОВКИ ИСХОДНЫХ ДАННЫХ ДЛЯ ПОСТРОЕНИЯ КРЕДИТНОГО СКОРИНГА

Т.И. Инхиреева
(г. Томск, Томский политехнический университет)
e-mail: tai2@tpu.ru

DATA PREPARATION METHODOLOGY FOR CREDIT SCORING

T.I. Inkhireeva
(Tomsk, Tomsk Polytechnic University)

Abstract. This paper considers data preparation methodology for logistic regression credit scoring model.

Keywords: data mining, data preprocessing, data cleaning, logistic regression, credit scoring.

Введение. Одной из главных задач, стоящих перед кредитно-финансовыми организациями, является оценка рисков выполнения заемщиком его кредитных обязательств. Анализ рисков предполагаемого заемщика производится на основе анкетных данных двумя способами – экспертной оценкой либо с помощью скоринговых систем [1].

Для построения скоринговой карты могут быть использованы различные методы прогнозирования: нейронные сети, деревья решений, логистическая регрессия и т.д.

Эффективность анализа во многом зависит от качества исходных данных, поэтому подготовка данных является очень важным этапом. Часто его игнорируют полностью или частично, что отрицательно отражается на результатах анализа. Полученные на этапе сбора данные обычно содержат недостатки: пропуски, дубли, недопустимые значения, невозможные комбинации значений и т.д. Данные могут иметь разный формат, обладать нежелательными для дальнейшего анализа свойствами (мультиколлинеарность, корреляция, распределение, отличное от нормального). Никакие методы не показывают хороших результатов на некачественных данных.

На практике чаще всего используется логистическая регрессия. Эта модель позволяет оценить вероятность возврата кредита для конкретного заемщика. В модели бинарной логистической регрессии целевая переменная $y \in \{0, 1\}$ отражает кредитоспособность заемщика.

Модель логистической регрессии. Математически модель логистической регрессии выражает зависимость логарифма шанса от линейной комбинации независимых переменных

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_{1,1} + ... + + b_k x_{i,j} + \varepsilon_i \quad (1)$$

где p_i— вероятность наступления дефолта по кредиту для i-го заемщика;
x_{ij} — значение j-ой независимой переменной;
b_0 — независимая константа модели, b_j — параметры модели;
ε_i — компонент случайной ошибки.

Уравнение (1) отражает линейную зависимость вероятности наступления просрочки по кредиту в зависимости от значений независимых переменных.

Постановка задачи. Требуется произвести подготовку данных для решения задачи бинарной классификации потенциальных заемщиков банка методом логистической регрессии. Исходными данными к работе являются исторические данные о кредитоспособности, содержащие числовые (возраст, доход), категориальные (гражданство, цель кредита), ранговые (уровень заработной платы) и бинарные переменные (наличие карты банка, наличие невыплаченных кредитов), одна из которых – целевая. Неплательщиками считаются заемщики, которые не платили по кредиту в течение 90 дней.

Целевая переменная является принимает значение 0, если заемщик не имеет просрочки (хороший) и 1, если заемщик имеет просрочку более 90 дней (плохой).

Разбиение данных. Для построения адекватной модели и проверки ее точности исходные исторические данные о заемщиках необходимо разбить на две или три независимые выборки, в зависимости от количества обучаемых моделей.

Если предполагается обучение одной предсказательной модели, рекомендуется разбиение на две выборки – тренировочную и тестовую. На тренировочной выборке происходит обучение моделей, то есть оптимизация их параметров. На тестовой выборке осуществляется оценка качества модели. Соотношение числа наблюдений в тренировочной выборке и тестовой чаще всего составляет 70-80% и 30-20% соответственно. Это соотношение определяется количеством исходных данных. Модель, обученная на небольшом количестве тестовых данных, обладает большой дисперсией, т. е. ее результаты на разных выборках сильно различаются. При среднем объеме выборки (тысячи или десятки тысяч наблюдений) соотношение 70% к 30% и 80% к 20% считается стандартным решением. Если количество наблюдений измеряется сотнями тысяч, есть смысл уменьшить тренировочную выборку и увеличить тестовую, особенно если предсказательная модель требует большого объема вычислений. При небольшом объеме исходных данных (сотни-тысячи наблюдений) стоит использовать кросс-валидацию [2].

В процессе кросс-валидации данные сначала разбиваются на тренировочную и тестовую выборку. Тренировочная выборка разбивается на k частей. Обучение происходит на k-1 частях, оставшаяся часть используется для валидации.

При обучении нескольких моделей используется разбиение на три выборки – тренировочную, валидационную и тестовую. Этим решается проблема переобучения. На валидационной выборке производится сравнение результатов работы нескольких моделей и выбор лучшей. Стандартным соотношением размера трех выборок является 60, 20 и 20% для данных среднего объема исходных данных. Данные выше рекомендации о выборе соотношения размера двух выборок применимы и в данном случае.

Вышеперечисленные схемы разбиения данных представлены на рис. 1.

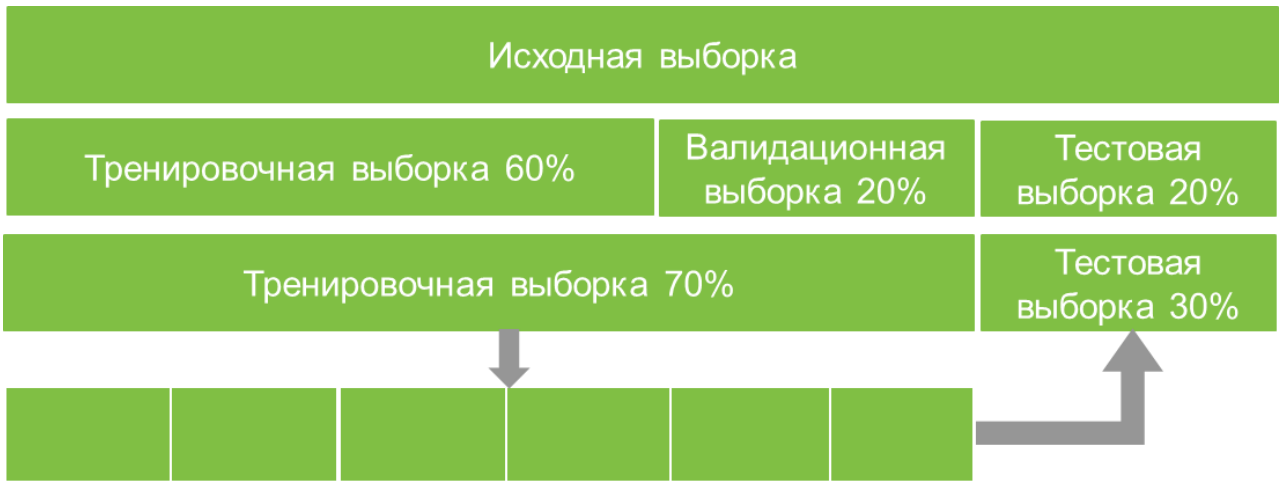

Рисунок 0 – Разбиение данных

При разбиении данных возникает проблема репрезентативности полученных выборок. Все выборки должны обладать тем же соотношением классов целевой переменной, что и в исходной выборке.

Очищение данных. Очищение данных включает в себя удаление дублирующихся записей, исправление ошибочных и противоречивых данных, обработку выбросов и пропущенных значений. Решение этих проблем может значительно улучшить качество прогнозной модели.

Дублирующиеся наблюдения. Наличие одинаковых наблюдений влияет на коэффициенты регрессии, увеличивая дисперсию модели, поэтому дублирующиеся наблюдения должны быть найдены и удалены из анализа [3].

Выбросы. Выбросы в данных – это аномальные значения, выделяющиеся из общей выборки. Логистическая регрессия чувствительна к выбросам, поэтому их обработка является очень важным шагом подготовки данных.

Простейший способ определения выбросов в числовой переменной – считать выбросом все наблюдения, которые не укладываются в заданные квантили. Графически этот подход реализован в виде диаграмм размаха (ящик с усами). Диаграмма размаха показывает медиану или среднее, межквартильный размах, максимальное и минимальное значение, выбросы.

Значительно труднее выявить многомерные выбросы. Двумерные выбросы можно выявить с помощью диаграммы рассеяния (точечной диаграммы). Построение точечных диаграмм для всех пар переменных помогает выявить двумерные выбросы. Наблюдения, являющиеся выбросами более высоких размерностей, можно выявить с помощью алгоритмов изолирующего леса и dbscan и многих алгоритмов кластеризации.

Выбросы в категориальных переменных могут быть обнаружены с помощью гистограмм.

При небольшом количестве выбросов можно удалить их из анализа или заменить средним, или модой. При большом количестве выбросов следует выделить их в отдельную выборку для проведения анализа, поскольку это может свидетельствовать о появлении нового феномена в данных. Помочь справиться с выбросами в числовой переменной может применение некоторых преобразований (min-max нормализация) и дискретизация [4].

Коррекция противоречивых данных. Неверные и противоречивые значения могут появиться на этапе ввода, передача и сбора данных в результате опечаток, программных ограничений (ограничение на длину переменной, ограничение размера буфера), различных форматов записи данных (Санкт-Петербург, Ст.-Петербург).

Неверными могут оказаться редкие категориальных переменных, экстремальные (рост: 250 см.) или необычные (заработная плата: -1 руб.) значения числовых переменных. Выявить такие значения можно с помощью гистограмм, ящика с усами или диаграмм рассе-

яния. Противоречивые данные (пол: мужской, беременность: да) можно выявить, используя релевантные логические правила. Для выявления некоторых ошибочных и противоречивых данных может понадобиться эксперт в предметной области [4].

Неверные и противоречивые данные представляют проблему потому, что алгоритм логистической регрессии предполагает, что все исходные данные – корректные, и строит модель в соответствие с этим предположением, что приводит к неверным результатам. При выявлении неверных или противоречивых значений, необходимо исправить или удалить соответствующее наблюдение из анализа.

Обработка пропусков. Пропуски в данных могут быть обусловлены множеством причин: необходимые данные не всегда могут быть доступны (информация о клиенте), данные могут отсутствовать потому, что считались не нужными в определенный момент времени. Пропуски также могут возникнуть из-за технических проблем. Данные могут быть удалены по причине противоречивости. Многие методы анализа данных, в том числе логистическая регрессия, не способны работать с пропущенными данными, поэтому пропуски необходимо тем или иным образом устранять: удалять наблюдения, содержащие пропуски либо заполнять их.

В случае необходимости заполнения поля на этапе ввода данных, пропущенные значения кодируют некоторым заменяющим значением, выбранным так, чтобы оно не было похожим на типичное для переменной значение.

В зависимости от причин, породивших пропущенные данные, пропуски могут иметь различное распределение. Понимание этого распределения может помочь выбрать алгоритм заполнения пропущенных данных. Механизмы появления пропущенных данных делятся на три категории:

Missing Completely at Random (MCAR) – механизм формирования пропусков, при котором вероятность пропуска для каждой записи одинакова. В таком случае игнорирование или исключение записей, содержащих пропущенные данные, не ведет к искажению результатов.

Missing at Random (MAR) – чаще всего данные пропущены не случайно, а ввиду некоторых закономерностей. Пропуски относят к MAR, если вероятность пропуска может быть определена на основе другой имеющейся в наборе данных информации (пол, возраст, занимаемая должность, образование), не содержащей пропуски. В таком случае удаление или замена пропусков на значение «Пропуск», как и в случае MCAR, не приведет к существенному искажению результатов.

Missing not at Random (MNAR) – механизм формирования пропусков, при котором данные отсутствуют в зависимости от неизвестных факторов. MNAR предполагает, что вероятность пропуска могла бы быть описана на основе других атрибутов, но информация по этим атрибутам в наборе данных отсутствует. Как следствие, вероятность пропуска невозможно выразить на основе информации, содержащейся в наборе данных.

На практике может быть не очевидно, к какой категории отнести пропущенные данные, потому что механизм их появления может быть просто неясен. Механизм MCAR может быть выявлен с помощью t-критерия Стьюдента или критерия Литтла [5]. Данные, содержащие менее 5% пропусков, можно считать MCAR. Для данных, содержащих от 5 до 50% пропусков, необходимо определить механизм их возникновения и в соответствие с этим выбирать стратегию их заполнения. Переменные, содержащие более 50% пропущенных значения, следует удалить из анализа. MAR и MNAR могут быть выявлены вручную, зачастую для этого требуется помощь эксперта в предметной области. Большая часть методов заполнения пропусков предполагает работу с данными MCAR и MAR, поскольку их присутствие не влияет существенным образом на результат.

Наиболее распространенными методами заполнения пропусков в числовых переменных являются: заполнение константой (нулем, средним, модой, медианой, последним наблюдением), заполнение из распределения, заполнение с помощью модели (нейросеть, дерево решений).

Для обработки отсутствующих значений в категориальных переменных используется, создание отдельной категории для пропущенных данных, создание бинарной переменной-индикатора.

Дискретизация. Дискретизация непрерывных переменных может быть предпочтительна, если распределение переменной мультимодально, имеет тяжелые хвосты, выбросы или пропущенные значения. В таких случаях дискретизованная версия непрерывной может упростить для анализа сложные нелинейные зависимости.

Нормализация (масштабирование). Принято считать, что масштабирование переменных не влияет на логистическую регрессию, однако неизбежное применение регуляризации l1 и l2 привносит необходимость масштабирования переменных перед использованием логистической регрессии.

Наиболее распространенные методы масштабирования – стандартизация и min-max нормализация. Стандартизация используется в случае приближенности распределения переменной к нормальному, в противном случае предпочтительна min-mix нормализация.

Нелинейное преобразование. Обычно для преобразования числовых переменных используют следующие виды преобразований: квадратное; кубическое; квадратный корень; натуральный или десятичный логарифм; экспоненциальное; величина, обратная квадратному корню; обратная величина. При использовании степенных преобразований ко всем значениям преобразуемой переменной могут добавлять константу для преобразования нуля или отрицательных значений. Такие преобразования количественных переменных могут привести к максимизации их связи с зависимой целевой переменной. Необходимое преобразование подбирается эмпирически, так чтобы полученная переменная наиболее точно описывала целевую переменную. Также часто используются относительные преобразования, например отношение суммы дохода к сумме задолженности [1].

Выбор переменных. Последним шагом перед непосредственным анализом данных является выбор переменных. На этапе выбора переменных отбрасываются неинформативные, избыточные переменные и переменные, которые не улучшают модель.

Мультиколлинеарность. Присутствие мультиколлинеарности в объясняющих переменных приводит к увеличению дисперсии модели логистической регрессии, получению неправильных знаков при оценке параметров модели, а также неустойчивости оценки параметров модели.

Для выявления мультиколлинеарности используется анализ корреляционной матрицы и статистика Variance Inflation Factor (VIF):

$$VIF = \frac{1}{1 - R_i^2},$$

где R_i – коэффициент детерминации регрессии i-й переменной на остальные объясняющие переменные.

Если показатель VIF больше пяти, это говорит о присутствии мультиколлинеарности. В этом случае необходимо удалить данную переменную из анализа либо использовать метод главных компонент для конструирования новых признаков вместо исходных;

Информативность. Предварительный анализ информативности объясняющих переменных, их влияния на целевую переменную помогает сократить количество рассматриваемых признаков.

Переменные, не имеющие взаимосвязи с целевой переменной (идентификационный номер клиента, фаза луны в день подачи заявки), должны быть удалены из анализа, также, как и переменные с дисперсией равной или близкой к нулю, что значит, что на исходной выборке она почти всегда принимает одно и то же значение.

Основными методами оценки информативности переменных являются критерий хи-квадрат и показатель информационного значения (IV).

Заключение. Банки принимают решение о выдаче кредита на основе анкетных данных заемщика. Кредитный скоринг автоматизирует этот процесс. Один из самых популяр-

ных методов кредитного скоринга – логистическая регрессия. Предложенная методология подготовки содержит необходимые шаги для повышения точности скоринга методом логистической регрессии.

ЛИТЕРАТУРА

1. Сергеевич С.А. Построение скоринговых карт с использованием модели логистической регрессии // Интернет-журнал Науковедение. – 2014. – Vol. 2.
2. Ng A. Machine learning yearning. 5th ed. – deeplearning.ai, 2018. – 116 p.
3. Anshu B. Data Preprocessing Techniques for Data Mining // Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets. New Delhi – 2011 – P. 6.
4. Abbott D. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. – Indianapolis: Wiley, 2014. – 427 p.
5. Little R.J.A. A Test of Missing Completely at Random for Multivariate Data with Missing Values // J. Am. Stat. Assoc. Taylor & Francis – 1988. – Vol. 83 – № 404. – P. 1198–1202.

МОДЕЛЬ ЭКОНОМИКИ ЗАМКНУТОГО ЦИКЛА В ЛЕСОПРОМЫШЛЕННОМ КОМПЛЕКСЕ РОССИИ. ИНСТИТУЦИОНАЛЬНЫЙ ПОДХОД

Б.О. Калюжный, Е.А. Монастырный

(Франция, аспирант, Томский политехнический университет) E-mail: borisk@tpu.ru

(Россия, г. Томск, д.э.н., профессор НИ ТПУ, профессор ТУСУР, заведующий лабораторией устойчивого развития социально-экономических систем, ТНЦ СО РАН) e.monastyrny@gmail.com

Научные руководители: профессор ШИП, Томский политехнический университет, д.э.н., Е.А. Монастырный,

профессор Университет Бургундии Франш-Конте, Франция, г. Дижон, PhD, К. Бомон

MODEL OF CLOSED-LOOP ECONOMY IN THE RUSSIAN TIMBER INDUSTRY. INSTITUTIONAL APPROACH

B.O. Kalyuzhny, E.A. Monastery

(Tomsk, PhD student, National Research Tomsk Polytechnic University) E-mail: borisk@tpu.ru

(Tomsk, Ph.D. in Economics, Professor at NI TPU, Professor at TUSUR, Head of the Laboratory of Sustainable Development of Socio-Economic Systems, TSC SB RAS) e.monastyrny@gmail.com

Abstract. Waste generation and increase in the share of waste in the timber industry seriously increases the risk of unsustainable use of economic resources and destruction of natural capital, the main component of sustainable development of the industry. The institutional approach will make it possible to assess the current situation and to propose adapted measures to maximize the rational exploitation of available resources and minimize the possibility of conflicts

Key words. Wastes in the forest-industrial complex, destruction of natural capital, sustainable development, closed-loop economy

Объектом анализа является лесопромышленный комплекс (ЛПК) России. Отходы от заготовки и переработки древесины представляют угрозу для устойчивого развития ЛПК России. В этом контексте интегрирование модели экономики замкнутого цикла (ЭЗЦ) рассматривается как инновационный путь для дальнейшего развития ЛПК России [Калюжный & Монастырный, 2019].

Актуальность данной работы заключается в использовании институционального подхода [North, 1994; Шаститко, 2002; Tirole, 2018] для анализа возможности интегрирования модели ЭЗЦ в стратегию развития ЛПК России. ЭЗЦ рассматривается как инструмент