УДК 519.68:336.77

57. О ПОДГОТОВКЕ ДАННЫХ ДЛЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Габриелян М.В., Хурс М.Д.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Шинкевич Е.А. – канд. физ.-мат. наук

Аннотация. В работе рассматриваются ключевые этапы обработки данных для машинного обучения, включая предобработку (устранение пропусков, аномалий и дубликатов) и подготовку (кодирование, масштабирование, анализ корреляций). На примере задачи кредитного скоринга демонстрируется влияние качественной обработки данных на эффективность моделей. Особое внимание уделено классификации пропущенных данных по Д. Рубину и методам их обработки, а также современным подходам к кодированию категориальных признаков.

Ключевые слова. Машинное обучение, предобработка данных, кредитный скоринг, пропущенные значения, кодирование признаков, масштабирование, корреляционный анализ, ROC-AUC, MICE, PhiK-корреляция.

Машинное обучение (Machine Learning, ML) — это раздел теории искусственного интеллекта, предметом которого является поиск методов решения задач путем обучения в процессе решения сходных задач.

Для построения таких методов используются средства алгебры, математической статистики, дискретной математики, теории оптимизации, численных методов, и других разделов математики.

Результат работы методов машинного обучения сильно зависит от данных. Это самый критически важный аспект, благодаря которому и возможно обучение алгоритма; именно поэтому машинное обучение стало столь популярным в последние годы. Но вне зависимости от терабайтов информации и экспертизы в Data Science, если не получается понять смысл записей данных, то машина будет практически бесполезной, а иногда и наносить вред.

Проблема в том, что во всех массивах данных есть изъяны. Если данные, используемые для прогнозирования неправильные, необъективные или неполные, то они могут привести к неточным результатам прогноза.

Предобработка и подготовка данных решают разные проблемы. Предобработка направлена на то, чтобы устранить аномалии данных, пропуски, выбросы и дубликаты, которые могли появиться из-за разных ошибок. Это базовый технический этап работы с данными.

Подготовка, в отличие от предобработки, должна преобразовать данные, чтобы все признаки стали понятными для модели. Этот этап специфичен для МО, другие специалисты его не проходят. Кроме того, подготовка данных для разных моделей может очень сильно различаться. Конкретные задачи — конкретные меры по подготовке под модель.

Предположим, что заказчик — кредитный отдел банка. Поэтому на примере одного датасета будет показано, как подготавливать данные, которые могут в дальнейшем использоваться при построении модели кредитного скоринга и в результате может быть выявлено, какие признаки в какой степени влияют на факт погашения кредита в срок.

Рассмотрим поэтапно, как происходит работа с данными на практике (рисунок 1).

Рисунок 1 – Этапы работы с данными

Этап 1. Для начала изучаем общую информацию о данных, проверяем, чтобы они соответствовали описанию.

Этап 2.1. Далее начинается предобработка данных. В реальных наборах данных пропущенные значения создают проблему для дальнейшей обработки, и не все алгоритмы машинного обучения умеют работать с данными, в которых есть пропуски. Большую ценность имеет подстановка или заполнение отсутствующих значений. Рассмотрим классификацию пропусков, которую в 1976 году предложил математик Дональд Рубин (Donald B. Rubin) [2]:

1. Полностью случайные пропуски (missing completely at random, MCAR) предполагают, что вероятность появления пропуска никак не связана с данными. Такие пропуски возникают, например, если измерительный прибор неисправен и случайным образом не записал часть наблюдений, или если один из образцов крови, изучаемых в лаборатории, оказался поврежден и по этой причине его характеристики выпали из исследования. Считается, что в реальности наблюдать полностью случайные пропущенные значения очень сложно. Какие-либо закономерности (то есть связи с наблюдаемыми или отсутствующими значениями) все равно существуют. Это приводит нас ко второй категории пропусков.

2. Случайные пропуски (missing at random, MAR) — вероятность появления пропуска зависит от некоторой известной нам переменной. Например, отсутствие ответа на определенный вопрос анкеты может зависеть от возраста респондента. Молодые охотнее отвечают на вопросы, люди более пожилого возраста скорее избегают ответа. Если мы знаем об этой особенности, то можем, правильно собирая и корректируя данные, добиться большей объективности.

3. Неслучайные пропуски (missing not at random, MNAR) — вероятность появления пропуска зависит, в том числе, от фактора, о котором мы ничего не знаем. Например, у весов может быть верхний предел измерения и любой образец выше этого предела автоматически не записывается. В опросах общественного мнения MNAR возникает, когда люди с более активной жизненной позицией (переменная, которую мы не измеряем) чаще дают ответы на вопросы интервьюера.

Далее рассмотрим стратегии работы с пропусками. По большому счету их две: *удаление и заполнение.* У каждого подхода есть свои достоинства и недостатки. Во многих случаях **удаление пропусков** (missing values deletion) может оказаться правильным решением, потому что в этом случае мы не «портим» данные. Удаление пропущенных значений хорошо работает (позволяет качественно обучить алгоритм), если мы считаем, что пропуски носят полностью случайный характер (MCAR). Единственным ограничением в этом случае будет достаточность данных для обучения после удаления пропусков. Удаление пропусков не всегда возможно. В этом случае прибегают к **заполнению пропусков** (missing values imputation) разными методами: одномерными или двухмерными.

Одномерные методы (single Imputation) — это заполнение с использованием данных одного столбца. Другими словами, чтобы заполнить пропуски мы берем данные того же признака. Данные в машинном обучении бывают двух видов: количественные и категориальные.

Самый простой способ работы с пропусками в *количественных данных* — заполнить пропуски константой. Например, нулем (подходит для алгоритмов, чувствительных к масштабу признаков). Заполнение константой позволяет не сокращать размер выборки, однако может внести системную ошибку в данные.

Для *категориальных признаков* в некоторых случаях можно провести дополнительное исследование.

Количественные данные можно заполнить средним арифметическим или медианой (statistical imputation).

У такого простого и понятного подхода тем не менее есть ряд недостатков: во-первых, когда в данных появляется большое количество одинаковых, близких к среднему, значений, мы снижаем ценную вариативность в данных. Кроме того, такое заполнение пропусков может быть некорректно: например, данные кредитного скоринга, где, если заполнить пропуски в столбце «Стаж» средним

значением или медианой, молодой сотрудник может получить больший стаж, чем у него есть на самом деле, а сотрудник в возрасте, меньший.

Заполнение внутригрупповым значением — более сложный способ заполнения пропусков количественного признака — вначале разбить данные на категории, а затем вычислить медианное значение для каждой категории и только потом заполнять им пропущенные значения.

Для заполнения пропусков в категориальных данных подойдет **метод заполнения наиболее часто встречающимся значением** (модой). Если пропусков немного, этот метод вполне обоснован. При большом количестве пропусков, можно попробовать создать на их основе новую категорию. В случае если у нас есть существенное количество пропусков в категориальной переменной, мы можем задуматься над созданием отдельной категории для пропущенных значений.

Многомерные методы (multivariate imputation) предполагают заполнение пропусков одной переменной на основе данных других признаков. Другими словами, мы строим модель машинного обучения для заполнения пропусков. Такой моделью может быть линейная регрессия для количественных признаков или логистическая — для категориальных.

Рассмотрим пример линейной регрессии и сразу два подхода к ее реализации, детерминированный и стохастический.

Детерминированный подход (deterministic approach) предполагает, что мы заполняем пропуски строго теми значениями, которые будут предсказаны линейной регрессией.

В детерминированном подходе сохраняется та же особенность, которую мы наблюдали при заполнении пропусков медианой, а именно доминирование одного значения (в случае медианы) или узкого диапазона (в случае линейной регрессии).

При применении **стохастического подхода** (stochastic regression imputation) добавляется гауссовский шум к предсказанным значениям, сохраняя вариативность данных и предотвращая излишнюю детерминированность.

MICE (Multiple Imputation by Chained Equations) – итеративно заполняет пропуски, предсказывая каждую переменную на основе других, пока процесс не стабилизируется. Это более сложный итеративный алгоритм, который строит модели для каждого признака, используя другие переменные.

Этап 2.2. Обработка аномалий (выбросов)

Выбросы могут искажать модель, поэтому их нужно выявлять и корректировать. Методами обнаружения являются визуальный анализ (боксплоты, гистограммы) и статистические методы (правило 3σ, межквартильный размах).

Этап 2.3. Поиск и устранение дубликатов

1. Явные дубликаты — полностью совпадающие строки.

2. Неявные дубликаты — строки, которые дублируют информацию, но с небольшими различиями (например, опечатки в категориальных данных).

Этап 3. Исследовательский анализ данных (EDA) — это ключевой этап, позволяющий глубоко понять структуру данных, выявить аномалии, пропуски и зависимости между признаками. Он включает в себя статистический анализ и визуализацию, что помогает принять обоснованные решения на этапах предобработки и построения модели [3].

Для начала требуется провести статистический анализ количественных признаков. Для каждого числового признака вычисляются основные статистические показатели: среднее, медиана, мода, стандартное отклонение, минимум, максимум, квартили (25%, 50%, 75%). Если среднее больше медианы, возможны выбросы в больших значениях, и, если стандартное отклонение велико, данные сильно разбросаны.

Для визуализации распределений количественных признаков используются гистограммы, которые показывают частоту значений в интервалах (рисунок 2). Также можно использовать boxplot (ящики с усами), на которых видно медиану (Q2), границы ящика (Q1, Q3), диапазон нормальных значений и выбросы. Если медиана не по центру ящика, распределение скошено. Длинные усы или много выбросов – данные имеют аномалии.

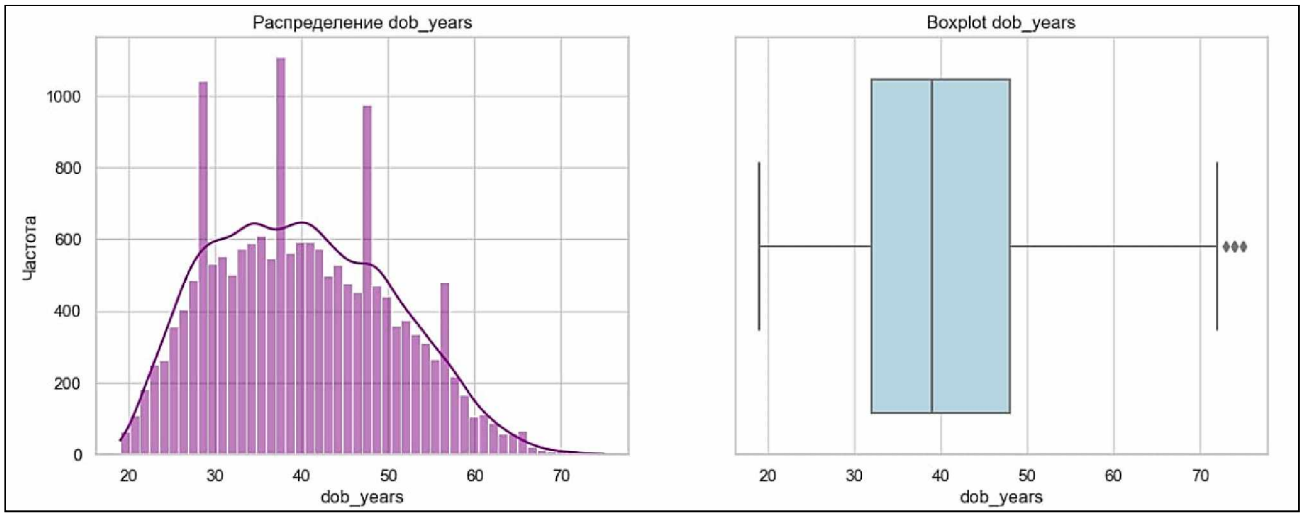

Рисунок 2 – Визуализация распределений количественных признаков

Для анализа качественных и дискретных признаков используются столбчатые диаграммы (bar plots), которые показывают частоту каждой категории; **круговые диаграммы (pie charts), которые** удобны для показа долей, но только при малом числе категорий.

Этап 4. Корреляционный анализ и проверка на мультиколлинеарность.

Корреляционный анализ — важный этап подготовки данных, который позволяет выявить взаимосвязи между признаками. Это помогает улучшить качество модели, исключив избыточные или слабо информативные признаки; обнаружить скрытые зависимости, которые могут влиять на целевую переменную; избежать мультиколлинеарности — ситуации, когда несколько признаков сильно коррелируют между собой, что ухудшает интерпретацию модели [4].

Лучший метод для выявления скрытых зависимостей – PhiK-корреляция (φ_k). Обычные методы (Пирсон, Спирмен) измеряют только линейную зависимость или работают с монотонными зависимостями, но не выявляют сложные линейные связи (плохо работают с категориальными данными). PhiK-корреляция работает с любым типами данными и обнаруживает нелинейные зависимости, которые не видны в классических методах. На практике строится матрица корреляции между всеми признаки, визуализируется в виде тепловой карты (heatmap) для наглядности, анализируются наиболее коррелирующие пары признаков.

Этап 5. Подготовка данных.

Существуют две основные процедуры подготовки данных: кодирование и масштабирование. Кодирование приводит категориальные признаки к понятным модели стандартам, так как большинство моделей не умеет напрямую работать с номинальными, порядковыми и бинарными данными в виде категориальных или текстовых форматах, поэтому их нужно закодировать как количественные. Существует два типа категориальных данных: номинальные (состоят из категорий без какого-либо внутреннего порядка или ранжирования) и порядковые (категории с определенным порядком и ранжированием, где важна взаимосвязь между значениями). Используя правильные методы кодирования можно эффективно повысить производительность и прогностические возможности модели. Рассмотрим виды кодирования данных [5].

1. Кодировка меток (Label Encoding). **Кодирование меток** присваивает каждой категории уникальное целое число. Однако оно не учитывает порядок категорий, что делает его более подходящим для **номинальных данных**, где порядок не имеет значения. Для **порядковых данных** лучше использовать **порядковое кодирование**, так как оно сохраняет порядковые отношения.

2. Одноразовое кодирование (One-Hot Encoding). **Однократное кодирование** преобразует категориальные данные в двоичный формат, где каждая категория представлена новым столбцом. Значение 1 указывает на наличие этой категории, а 0 — на её отсутствие. Этот метод идеально подходит для **номинальных данных** (категорий без порядка), не позволяя модели предполагать наличие каких-либо связей между категориями.

3. Порядковая кодировка (Ordinal Encoding). **Порядковое кодирование** используется для **порядковых данных**, где категории имеют естественный порядок. Оно преобразует категориальные значения в числовые, сохраняя присущий им порядок.

4. Целевое кодирование (Target Encoding). **Целевое кодирование** (также известное как кодирование средним значением) — это метод, при котором каждая категория признака заменяется

средним значением целевой переменной для этой категории. Этот метод особенно полезен, когда существует связь между категориальным признаком и целевой переменной.

5. Двоичное кодирование (Binary Encoding). **Бинарное кодирование —** это более компактная версия однократного кодирования. Каждой категории присваивается уникальный двоичный код. Затем двоичный код разбивается на несколько столбцов. Этот метод подходит для наборов данных с высокой кардинальностью (множеством уникальных категорий), так как в результате получается меньше столбцов по сравнению с однократным кодированием.

6. Частотное кодирование (Frequency Encoding). **Частотное кодирование** присваивает каждой категории значение, основанное на ее частоте в наборе данных. Этот метод может быть полезен для обработки категориальных признаков с высокой кратностью (признаков с множеством уникальных категорий).

Если количественные признаки слишком сильно отличаются масштабом, то нужно привести их к одному масштабу [6]. *Масштабирование* изменяет данные без искажений их структуры и сохраняет все зависимости между признаками при переводе в новый диапазон значений.

В машинном обучении есть два распространенных способа масштабирования функций: нормализация и стандартизация. MinMaxScaler для нормализации и StandardScaler для стандартизации. Разница между этими двумя методами заключается в том, что нормализация изменяет масштаб данных, так что в итоге мы получаем значения от 0 до 1, а стандартизация изменяет масштаб данных, так что среднее значение становится равным 0, а стандартное отклонение становится равным 1.

После всей проделанной предобработки и подготовки данных можно использовать пайплайны, обучать модели, подбирать гиперпараметры и анализировать результаты.

В ходе практического исследования с помощью пайплайна без предобработки и подготовки данных было обучено несколько моделей классификации и выбрана лучшая модель с метрикой ROC-AUC на тестовой выборке, равной 0,4989, что является фактически случайным угадыванием. А после выполнения всех вышеописанных шагов метрика ROC-AUC стала равной 0,6097, что является реальным улучшением на 11,08 процентных пунктов. Это значимое улучшение метрики, так как модель перешла от "не лучше случайного" к "имеющей предсказательную силу". Это означает, что данные действительно требовали предобработки и подготовки для того, чтобы модель научилась использовать входную информацию эффективно.

Список использованных источников:

1. Введение в машинное обучение. [Электронный ресурс]. – Режим доступа: https://habr.com/ru/articles/448892/. – Дата доступа: 24.02.2025.

2. Пропущенные значения. [Электронный ресурс]. – Режим доступа: https://www.dmitrymakarov.ru/data/missing/. – Дата доступа: 28.02.2025.

3. Исследовательский анализ данных. [Электронный ресурс]. – Режим доступа: https://sky.pro/media/kak-provodit-issledovatelskij-analiz-dannyh/. – Дата доступа: 28.02.2025.

4. Phik Correlation Constant. [Электронный ресурс]. – Режим доступа: https://phik.readthedocs.io/en/latest/. – Дата доступа: 03.03.2025.

5. Categorical Data Encoding Techniques in Machine Learning. [Электронный ресурс]. – Режим доступа: https://www.geeksforgeeks.org/categorical-data-encoding-techniques-in-machine-learning/. – Дата доступа: 05.03.2025.

6. Feature Scaling in Machine Learning. [Электронный ресурс]. – Режим доступа: https://medium.com/codex/feature-scaling-in-machine-learning-e86b360d1c31. – Дата доступа: 12.03.2025.

UDC 519.68:336.77

ON DATA PREPARATION FOR MACHINE LEARNING

Khurs M.D., Gabrielyan M.V.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Shinkevich E.A. – PhD in Physics and Mathematics

Annotation. The paper examines the key stages of data processing for machine learning, including preprocessing (handling missing values, anomalies, and duplicates) and feature preparation (encoding, scaling, correlation analysis). Using credit scoring as an example, the study demonstrates how proper data processing impacts model performance. Special attention is given to Rubin's classification of missing data and processing methods, as well as modern approaches to categorical feature encoding.

Keywords. Machine learning, data preprocessing, credit scoring, missing values, feature encoding, scaling, correlation analysis, ROC-AUC, MICE, PhiK correlation.