МЕТОДИЧЕСКИЕ ЭКСПЕРИМЕНТЫ

С.В. Жучкова, А.Н. Ротмистров
(*Москва*)

ВОЗМОЖНОСТЬ РАБОТЫ С ПРОПУЩЕННЫМИ ДАННЫМИ ПРИ ИСПОЛЬЗОВАНИИ CHAID: РЕЗУЛЬТАТЫ СТАТИСТИЧЕСКОГО ЭКСПЕРИМЕНТА[1]

Рассматривается вариант работы с пропущенными данными («пропусками») «как есть», т.е. предполагающий придание пропускам статуса самостоятельной категории изучаемой переменной. Этот вариант кардинально отличается от других вариантов работы с пропусками: удалять те наблюдения, которые содержат пропуски, или заполнять пропуски. Один из известных нам методов, позволяющий реализовать вариант работы с пропусками «как есть» – CHAID. Модели деревьев с пропусками нередко встречаются в эмпирических исследованиях, однако в литературе отсутствует систематическое рассмотрение вопроса, какие конкретно преимущества и ограничения имеет реализованный в CHAID вариант работы с пропусками «как есть» по сравнению с обозначенными альтернативными

Светлана Васильевна Жучкова – студентка магистратуры факультета компьютерных наук, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: lana_lob@mail.ru.
Алексей Николаевич Ротмистров – кандидат социологических наук, доцент кафедры методов сбора и анализа социологической информации, департамент социологии, факультет социальных наук, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: alexey.n.rotmistrov@gmail.com.

[1] Публикация подготовлена в ходе проведения исследования «Обоснование преимуществ поиска эффектов взаимодействия и их учета в социологических регрессионных моделях» (№18-05-0031) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2018 г. и в рамках государственной поддержки ведущих университетов Российской Федерации «5–100».

вариантами. С целью начать дискуссию по этому поводу, мы провели несколько серий статистических экспериментов на модельных данных, организованных в три переменные категориального и интервального типа. Было эмпирически установлено, что в целом метод корректно распределяет пропуски по узлам, однако в большинстве случаев включение пропусков в анализ сопровождается изменениями в структуре дерева, а следовательно, существует риск получения неверных, ложных, ошибочных выводов. Также представлены рекомендации о том, какие факторы следует учитывать при принятии решения о включении пропусков в модель «как есть».

Ключевые слова: деревья решений, деревья классификации, категориальные переменные, поиск взаимодействий, пропущенные данные, пропущенные значения, статистический эксперимент, CHAID.

Введение

Проблема наличия пропусков в данных актуальна для социолога и часто становится препятствием к применению различных методов анализа, так как большинство из них требуют либо заполнения (импутации) пропусков, либо удаления наблюдений, содержащих пропуски. Методы заполнения пропусков хотя и пользуются популярностью среди социологов-практиков, однако часто подвергаются критике. Отмечаются следующие ограничения: во-первых, все методы заполнения требуют случайного (missing at random) или полностью случайного (missing completely at random) характера пропусков [1], и это ограничение одновременно легко нарушить и тяжело проверить [2, p. 270]; во-вторых, методы заполнения либо не учитывают нелинейный характер связи между признаками [3, p. 92], либо требуют для этого дополнительной трансформации данных [4]; в-третьих, использование искусственно созданных наблюдений опасно тем, что создает «иллюзию полноты данных, может смещать исходную структуру связей и приводить к неверным выводам» [5, p. 3].

«Единственное действительно хорошее решение проблемы пропущенных данных – вообще их не иметь» [6, p. 2] – так до сих пор звучит наиболее универсальный совет относительно работы с пропусками, который на практике почти не реализуем. В связи с этими обстоятельствами встает вопрос: можно ли работать с пропусками «как есть»? Такой вариант реализован во многих методах построения деревьев решений. Нами рассматривается здесь наиболее распространенный в социологии метод построения деревьев – CHAID. Широту функционала этого метода и пользу его применения в социологии сложно переоценить: он используется в задачах регрессии (предсказания количественного отклика), классификации (предсказания категориального отклика), кластеризации или сегментирования, поиска взаимодействий. По мнению Л. Рокаха и О. Меймона, одновременно с широкими возможностями метод обладает и конкретными преимуществами: универсальностью (так как приспособлен к анализу переменных любого типа шкалы и не имеет ограничений на параметры и форму распределения), доступностью (так как представлен во всех статистических пакетах), простотой реализации и легкостью интерпретации результатов [7, p. 183–184].

Возможность работы с пропущенными значениями выделяется в литературе как уникальное преимущество деревьев решений [8, p. 297]. Принципы такой работы различаются в зависимости от метода построения дерева. Так, в популярном у социологов методе CHAID пропущенные значения рассматриваются как единая категория независимой переменной (далее – предиктора), которая при построении модели присоединяется к наиболее похожему по распределению отклика узлу [9] (узел – сочетание значений предикторов, определяющее некоторое распределение отклика). Если предиктор, содержащий пропуски, оказывается значимым в итоговой модели, в отдельных узлах наряду с валидными категориями появляется и категория «пропуск», содержащая в себе множество неизвестных значений (*рис*. 1 с фрагментом реального эмпирического исследования).

Одновременно с выделением этого преимущества метода до сих пор не существует доказательств корректности включения пропусков в модели «как есть», как и практически не существует работ, анализирующих различные варианты обработки пропусков в деревьях. Об этом свидетельствует как отсутствие соответствующих публикаций в зарубежных базах данных научного цитирования (исключение составляет работа [10], где рассматривается не CHAID, а другой метод построения деревьев решений – CRT), так и прямое указание на эту проблему в других источниках [11, p. 864].

Между тем, понимание, как именно реализуется включение категории «пропуски» в модели и к чему это приводит, становится особенно важным в связи с ростом популярности соответствующих методов[1]. Если включение реализуется некорректно, то социологи рискуют – как и в случае с заполнением (импутацией) данных – получать ложные выводы, если корректно, – то модели деревьев решений могут по праву занять одно из ключевых мест в арсенале методов социолога, предоставив не только широкий функционал, но и уникальную возможность работы с пропусками.

Модели деревьев решений и возможность работы с пропущенными данными

Существующие модели деревьев решений можно разделить на три семейства – CART, ML и AID, которые различаются областью применения и используемым критерием расщепления узлов. Модели семейства CART и ML работают с различными информационными критериями, показывающими степень неоднородности получаемых узлов, но модели первого семейства

[1] Так, при запросе «decision tree» OR «classification tree» с ограничением отрасли знаний до Social Sciences в базе данных научного цитирования Scopus обнаружится более 2000 публикаций, причем почти половина из них появилась за последние 5 лет.

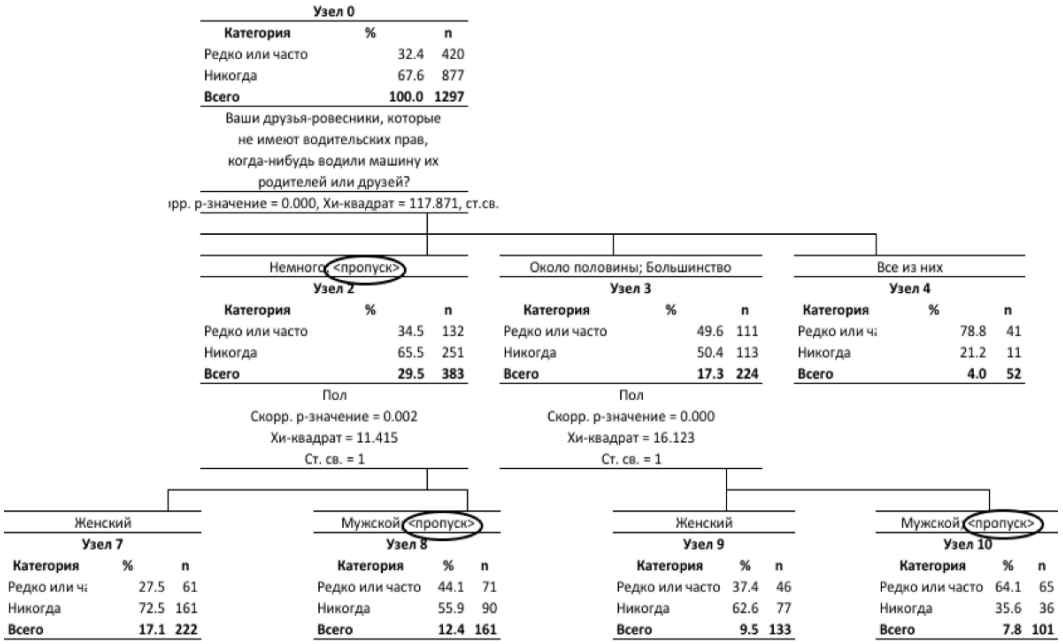

Рис. 1. Пример дерева, узлы которого содержат пропущенные значения

Источник: [12].

имеют свои истоки в статистике, а модели второго семейства – в компьютерных науках [11, p. 834–835]. Семейство моделей AID появилось в социальных науках для применения их к опросным данным [13, p. 49], и в этих моделях использовались разные критерии расщепления. В CHAID исходно применялась величина хи-квадрата Пирсона [11, p. 835]. Использование этой статистики в оригинальной модели было обусловлено номинальной шкалой зависимой переменной (далее – отклика) [9], однако в современных моделях, которые не имеют ограничений на тип шкалы отклика, выбор статистики более широк: для откликов интервального типа и выше применяется *F*-тест, для порядковых – тест отношения правдоподобия, для номинальных – хи-квадрат Пирсона [7, p. 181–182], а критерием расщепления выступает *p*-значение, полученная в соответствующих тестах. Различие в критериях расщепления приводит и к другому важному отличию выделенных семейств моделей: основная задача алгоритмов семейства AID – выделение групп, максимально отличающихся между собой по распределению отклика (максимизируется вариация отклика между группами, т.е. связь между предикторами и откликом), задача же алгоритмов семейств ML и CART – выделение максимально гомогенных групп (минимизируется вариация отклика внутри групп) [13, p. 53].

Исходная цель введения алгоритмов семейства AID в анализ социологических данных была вполне конкретной: они использовались не столько для улучшения качества прогноза, сколько для поиска значимых взаимодействий признаков [13, p. 51].

Алгоритм построения дерева CHAID включает несколько шагов: перед разбиением рассчитываются величины хи-квадрата (или иной статистики) между откликом и каждым предиктором; предиктор из пары с наименьшей величиной значимости выбирается первым из разделяющих; затем попарно рассматриваются распределения отклика при фиксации каждой из категорий выбранного предиктора; для каждой пары рассчитывается статистика хи-квадрат, и категории из пар с наибольшей величиной значимо-

сти объединяются в одну – этот шаг повторяется до тех пор, пока соответствующие величины значимости не достигнут заранее заданного порогового уровня; после этого объединенные категории, содержащие более двух исходных значений, «перепроверяются»: ищется возможный вариант бинарного статистически значимого разделения, если такого варианта не находится, формируется узел первого уровня глубины; описанная процедура повторяется со следующими предикторами [9, p. 121]. При этом для каждого предиктора учитывается тип шкалы, который определяет возможный порядок объединения категорий.

В отличие от других методов построения деревьев, CHAID обладает дополнительными техническими преимуществами. Во-первых, по мнению Дж. Ритчарда, этот метод не ограничен требованиями на число расщепляемых узлов и сам ищет их оптимальное число. Большинство же методов построения деревьев осуществляют либо бинарное расщепление, либо расщепление на число узлов, равное числу категорий отклика [13]. Во-вторых, как уже было отмечено, в отличие от более ранних моделей деревьев решений, благодаря возможности выбора статистики CHAID приспособлен к работе с откликами любого типа шкалы. В-третьих, благодаря встроенной опции поправки на множественные сравнения, этот метод более устойчив к статистическим ошибкам 1-го рода. Наряду с преимуществами CHAID отмечаются и недостатки, характерные для всех методов построения деревьев: чувствительность к изменениям в данных, опасность переобученности модели, неэффективность при малом объеме выборки [13, p. 77]. Однако эти ограничения, согласно [7], постепенно преодолеваются благодаря активному усовершенствованию моделей: их устойчивость (в вопросах как о размере выборки, так и чувствительности к изменениям в данных) повышается за счет использования ансамблей деревьев вместо одиночных моделей [14], а переобученность тестируется с помощью процедуры кросс-валидации или разделения выборки на обучающую и тестовую подвыборки.

Как уже было отмечено, при работе CHAID с пропущенными значениями их замены на действительные не происходит; пропуски рассматриваются как единая категория предиктора, которая при построении модели присоединяется к наиболее похожему по распределению отклика узлу – по такому же принципу, что и валидные категории. Предиктор, содержащий пропуски, имеет специальное название – «плавающий» (“floating”); «плавающим» он становится потому, что для одной из его категорий не определен тип шкалы и эта категория может присоединиться к любым другим категориям этого предиктора [9, p. 122]. Именно указанием возможности работы с пропусками и определением плавающего предиктора обычно ограничивается описание этого уникального преимущества CHAID в литературе. Следует отметить, что алгоритмы деревьев решений применяются и в случае, когда пропуски наблюдаются в отклике (как одни из методов заполнения пропусков [2, p. 272]), однако такой вариант применения алгоритма не рассматривается в настоящем исследовании.

Методология исследования

Цель исследования – установить, насколько корректным можно считать определение пропущенных значений по узлам дерева при использовании CHAID и какими последствиями сопровождается включение пропусков в модель. Для поиска ответов на поставленные вопросы был реализован статистический эксперимент. Идея эксперимента состояла в следующем: сформировать структуру многомерных связей между категориальным откликом и парой категориальных и одной интервальной переменной, эту структуру реализовать в сгенерированной модельной базе данных, построит на этой базе модель дерева, которая, очевидно, будет выражать заложенную исходно структуру многомерных связей, после чего в базе многократно случайным образом заменить часть данных на пропущенные значения и зафиксировать изменения в

структуре дерева, которые появляются при внесении пропусков. Сравнивая новые деревья с исходным, возможно определить, насколько корректно пропуски определяются по категориям. Иными словами, если исходные значения данных и их место в структуре дерева известны, то при целенаправленной замене этих значений на пропуски есть возможность определить правильность их отнесения к определенным узлам. Процедура эксперимента включала пять основных этапов (*рис*. 2).

Для анализа были выбраны предикторы номинального типа. Это связано с привычными областями применения CHAID: в социологии этот метод особенно полезен для поиска взаимодействий категориальных предикторов. В итоговую модель дерева было включено три предиктора – два номинальных (на глубине 1 и глубине 3) и интервальный (на глубине 2). Тип шкалы промежуточной переменной не имеет значения, поскольку в анализе эта переменная не участвовала, однако ее введение позволило рассмотреть ситуации, когда исследуемые переменные играют разные роли в структуре дерева, поскольку располагаются на минимально и максимально возможной глубине. Отклик был выбран дихотомический, чтобы облегчить процедуру проведения эксперимента, анализа результатов и генерирования базы данных.

Результаты экспериментов рассматривались в разрезе следующих условий: положение переменной в структуре дерева (глубина, отражающая силу связи), точность исходной модели дерева и доля пропусков в данных. Модель дерева, созданная для эксперимента и соответствующая идеальной ситуации, когда верно предсказываются 100% наблюдений, представлена на *рис*. 3. На *рис*. 4 и 5 представлены модели деревьев с точностью предсказаний 75% (для переменных на разной глубине). Модели деревьев, с которыми проводилось сравнение в ходе экспериментов, далее называются *исходными деревьями*.

Под представленные модели дерева с помощью логистической регрессии были сгенерированы три базы данных объемом в 3000 на-

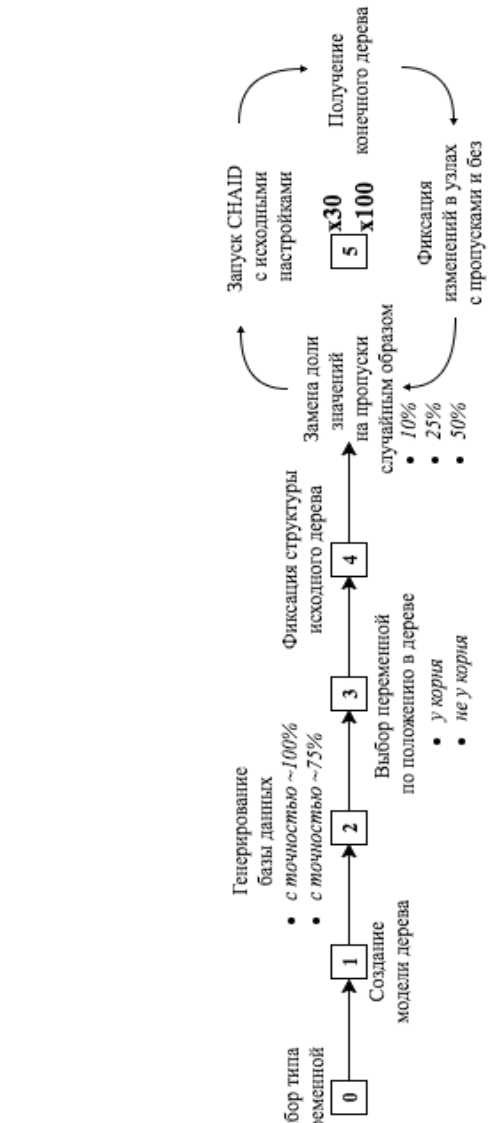

Рис. 2. Схема эксперимента

блюдений (каждая база данных воспроизводила связи, заложенные в моделях). Технические детали процесса генерирования базы данных для проведения экспериментов подробно описаны в *Приложении* 1.

Во всех моделях были зафиксированы структура узлов с номинальными переменными, т.е. было определено такое положение и содержание узлов, которое рассматривалось как точно воспроизводящее структуру связей в данных и с которым затем происходило сравнение. Например, в дереве с точностью 100% для переменной на глубине 3 фиксировалось, что один крайний узел объединяет в себе категории «1» и «2» (в этом узле 100% «единиц» по отклику), другой крайний узел – категории «3» и «4» (и в этом узле 100% «нулей» по отклику) – см. *рис*. 3.

Затем часть данных в независимой переменной заменялась случайным образом на пропущенные значения; эксперименты проводились с долями пропусков 10, 25 и 50% (далее – низкая, средняя и высокая доли пропусков). После этого CHAID запускался для обновленных данных, но с исходными настройками, чтобы обеспечить возможность сравнения моделей. После получения нового (далее – *конечного*) дерева фиксировались его точность, устойчивость (через кросс-валидацию с разбиением выборки на 10 подвыборок), а также анализировались изменения в узлах с пропусками и без. Этот основной, пятый этап эксперимента проводился в статистическом пакете IBM SPSS Statistics (подробнее см. *Приложение* 2 с фрагментом syntax).

Поскольку в CHAID пропущенные значения как отдельная категория присоединяются к тому узлу, который наиболее похож на эту категорию по распределению отклика, правило определения корректности отнесения пропущенных значений было следующим: *если узел, к которому были присоединены пропуски, по своей структуре и расположению не отличался от аналогичного узла в исходном дереве, то считалось, что пропуски были присоединены корректно и новый узел отражал действительную структуру связей в данных.*

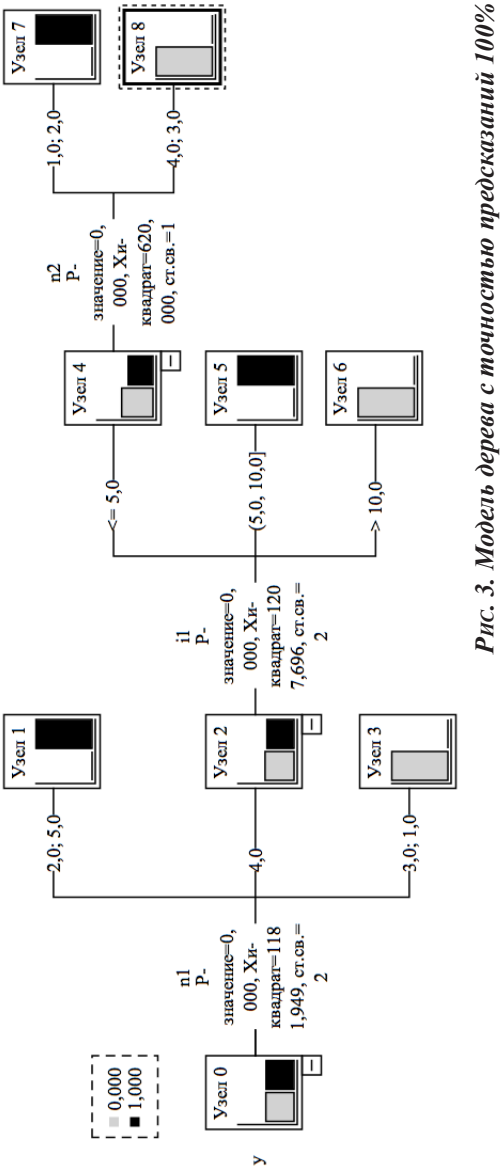

Рис. 3. Модель дерева с точностью предсказаний 100%

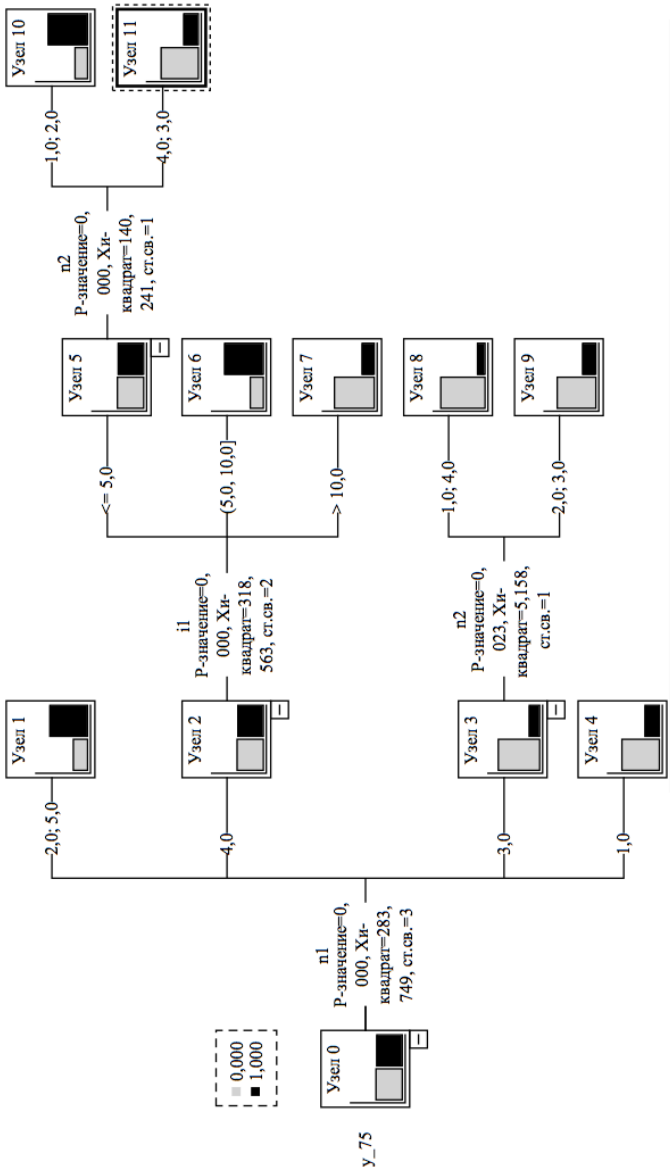

Рис. 4. Модель дерева с точностью предсказаний 75% (для переменной на глубине 1)

Согласно этому правилу, корректным считалось присоединение пропусков, если вновь получившийся узел соответствовал одновременно трем критериям.

1. Располагался на той же глубине, что и соответствующий ему узел исходного дерева.

2. Категории предиктора, с которыми были объединены пропуски, были идентичны категориям в соответствующем узле исходного дерева.

3. Распределение отклика в этом узле статистически значимо не отличалось от распределения в соответствующем узле исходного дерева (отсутствовали статистически значимые различия между долями «единиц» и «нулей» по отклику в конечном и исходном деревьях).

Если же при присоединении пропусков к узлу какой-либо из пунктов не соблюдался, то считалось, что пропуски были присоединены неверно. Исключение составляли случаи экспериментов с высокоточными деревьями (с точностью предсказаний 100%): в них корректным считалось отнесение пропусков к отдельному узлу, не существующему в исходном дереве. Дело в том, что наблюдения, замененные на пропуски, как подвыборка воспроизводят распределение всей выборки (в нашем случае – равномерное), а значит, содержат как нулевую, так и единичную категории отклика. И поскольку в высокоточных деревьях конечные узлы содержали по 100% нулевой или единичной категории, корректным считалось создание из пропусков отдельного узла, а не их присоединение к «чистым» узлам.

Кроме перечисленных параметров фиксировались также и другие ситуации, которые рассматривались как случаи *порчи дерева*, когда изменялась структура исходного дерева.

Описанный цикл (от замены значений на пропуски до фиксации изменений в структуре дерева) для каждого из сочетаний условий повторялся 30 или 100 раз (для экспериментов с точностью исходного дерева 100% – 30 раз из-за низкой вариации результатов). Множество экспериментов, относящихся к одной

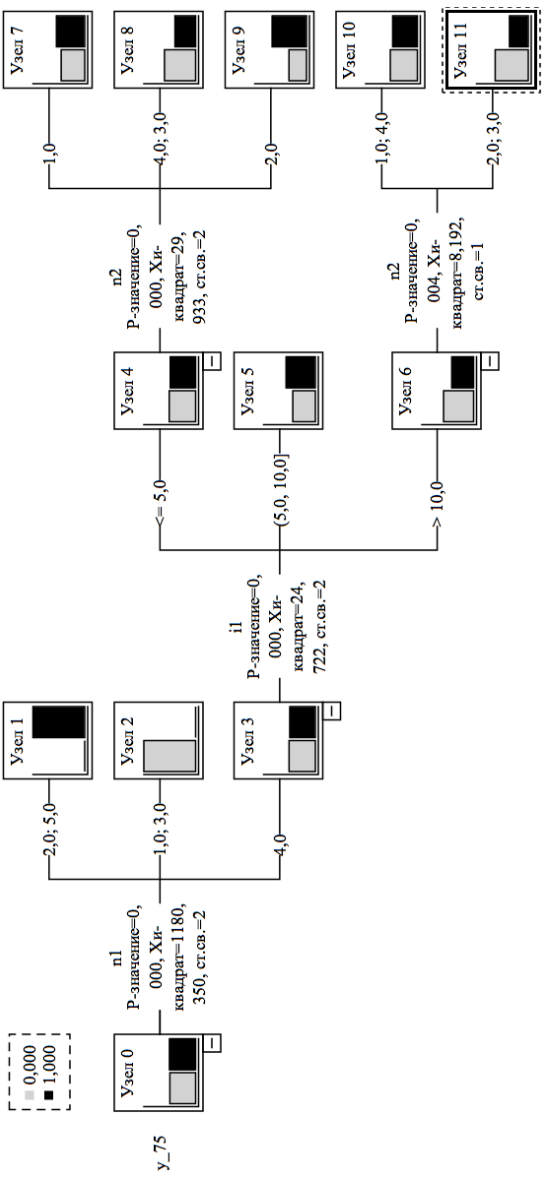

Рис. 5. Модель дерева с точностью предсказаний 75% (для переменной на глубине 3)

Таблица 1

РЕЗЮМЕ ПО ЧИСЛУ ПРОВЕДЕННЫХ ЭКСПЕРИМЕНТОВ

		Положение переменной				Всего экспериментов
		не у корня (глубина 3)		у корня (глубина 1)		
		Точность исходного дерева, %				
		100	75	100	75	
Доля пропусков	10	30	100	30	100	260
	25	30	100	30	100	260
	50	30	100	30	100	260
Всего экспериментов		390		390		780

из комбинаций каждого из трех условий, далее называется *серией экспериментов*. Так, всего было проведено 780 экспериментов, или 12 серий экспериментов (*табл.* 1). Эксперименты дали обширные результаты, поэтому дальнейшее их описание мы скомпоновали в несколько разделов.

Результаты экспериментов: верность определения пропусков

Как было отмечено ранее, в исследовании преследовались две цели: определить, насколько корректным можно считать отнесение пропусков по узлам алгоритмом CHAID, и установить, к каким последствиям приводит включение пропусков в анализ. Логика описания результатов экспериментов также строится в этих двух направлениях.

Гипотезами предварялось первое направление исследования – насколько корректно пропуски распределяются по узлам. *Основная гипотеза: в целом (независимо от условий экспериментов) пропуски определяются по узлам неверно, т.е. доля случаев с неверно определенными пропусками статистически значимо превышает долю случаев, когда они определены верно*. Такая

гипотеза обосновывается отсутствием в литературе каких-либо доказательств корректности работы CHAID с пропущенными данными. *Вспомогательные гипотезы строились на основе двух условий экспериментов: предполагалось, что точность определения пропусков 1) ниже в случае переменной не у корня по сравнению с переменной рядом с корнем, 2) снижается с увеличением доли пропущенных значений.*

Все статистические выводы далее сделаны с помощью расчета доверительных интервалов на уровне значимости $\alpha = 0{,}05$. Основная гипотеза не подтвердилась: в целом (без учета специфики серий экспериментов) пропуски распределились по категориям верно в 71% экспериментов (доверительный интервал составляет 68–75%, т.е. больше половины случаев). Это означает, что в CHAID узлы, к которым присоединяются пропуски, а) статистически чаще располагаются на «корректной» глубине, чем на «некорректной», и б) объединяют «верные» категории предиктора и содержат «правильное» распределение отклика, т.е. воспроизводят действительную структуру связей, заложенных в данных. Иными словами, узлы с пропусками в 68–75% случаев (на уровне значимости $\alpha = 0{,}05$) повторяют узлы, которые получились бы, если бы на месте пропусков исследователь имел их исходные значения. Прилагательные заключены в кавычки лишь потому, что в реальной практике исследователю неизвестна структура исходного дерева.

Была отвергнута и первая дополнительная гипотеза, где верность определения пропусков по узлам связывалась с расположением переменной по отношению к корню дерева. Так, для переменной, располагающейся рядом с корнем, пропуски были определены верно в 72% экспериментов (68–77%); для переменной, располагающейся «вдали» от корня, – в 71% экспериментов (66–75%). Пропуски определяются по узлам корректно независимо от положения переменной с пропусками в дереве.

Значительные различия в точности определения пропусков по узлам наблюдались в экспериментах с разной долей пропусков,

однако эти различия дают основания подтвердить вторую дополнительную гипотезу лишь частично. Так, для экспериментов, где доля пропущенных значений составляла 10%, пропуски определялись корректно в 91% случаев (88–95%), для экспериментов с долей пропусков 25% – в 87% случаев (82–91%), а для экспериментов с долей пропусков 50% – лишь в 37% случаев (31–42%). Таким образом, для низкой и средней доли пропусков присоединение их к узлам происходит одинаково корректно (в результатах для этих экспериментов нет статистически значимых различий между долями). Несмотря на то что случайные пропуски повторяют исходное распределение переменных, именно эксперименты с высокой долей пропусков (50%) отличаются наихудшими результатами, которые уменьшают и значение доли верных определений по всем экспериментам. *Подобные результаты дают повод усомниться в адекватности проведения анализа в случае, когда половину данных составляют пропущенные значения – вопреки тому, что CHAID не содержит ограничений на долю пропусков и способен составить решение (зачастую – высокоточное) при любых обстоятельствах. Результаты показывают, что полученное при высокой доле пропусков решение чаще, чем в половине случаев не соответствует заложенной структуре связей в данных. Таким образом, именно высокая доля пропусков может стать препятствием к включению пропущенных значений в анализ методом CHAID.* Более подробные результаты в разрезе каждой серии экспериментов представлены в *табл.* 2.

При идеальной точности исходной модели дерева пропуски определяются по узлам корректно всегда – в зависимости от распределения отклика в существующих узлах они будут присоединяться либо к максимально похожему по этому критерию узлу, либо формировать отдельный узел, если подходящего в дереве не заложено. При меньшей точности предсказаний исходного дерева результаты варьируются в зависимости от положения переменной по отношению к корню. Однако в обоих случаях – для переменной на глубине 1 и на глубине 3 – результаты статистически значимо

Таблица 2

ДОЛЯ ЭКСПЕРИМЕНТОВ С ВЕРНО ОПРЕДЕЛЕННЫМИ ПРОПУСКАМИ

		Положение предиктора											
		не у корня (глубина 3)						у корня (глубина 1)					
		Точность исходного дерева, %											
		75			100			75			100		
		Доля	Н. гр. ДИ	В. гр. ДИ	Доля	Н. гр. ДИ	В. гр. ДИ	Доля	Н. гр. ДИ	В. гр. ДИ	Доля	Н. гр. ДИ	В. гр. ДИ
Доля пропусков	10	0,77	0,69	0,85	1	1	1	1	1	1	1	1	1
	25	0,74	0,65	0,83	1	1	1	0,91	0,85	0,97	1	1	1
	50	0,35	0,26	0,44	1	1	1	0	0	0	1	1	1

Примечание. Н. гр. ДИ – нижняя граница доверительного интервала; В. гр. ДИ – верхняя граница доверительного интервала.

не отличаются для низкой и средней доли пропусков, а в случае с высокой долей пропусков большая часть решений оказывается неверной. Причем в ситуации, когда для исходного дерева с точностью 75% доля пропусков составляет 50% по переменной рядом с корнем, алгоритм CHAID вообще перестает определять их правильно. Это связано с изменением глубины расположения переменных, которое будет подробно описано далее.

Если исключить из сравнения случаи с высокой долей пропусков, то обнаружится, что при долях 10 и 25% с задачей верного определения пропусков по узлам CHAID лучше справляется для переменной, расположенной рядом с корнем. Причем и для низкой, и для средней доли пропусков вероятность правильного отнесения очень высока: 100% для низкой доли пропусков и 85–97% – для средней. Обратные результаты обнаруживаются при высокой доле пропусков: задача верного определения пропущенных значений по узлам успешнее решается для переменной не у корня – хотя и в этом случае доля неверных определений превышает долю верных.

Как итог первой части анализа результатов можно заключить, что в целом CHAID корректно определяет пропуски по узлам, а наихудшие результаты в присоединении их к категориям наблюдаются лишь в случае, когда доля пропущенных значений очень велика (50%). Однако речь до сих пор шла лишь о том, что происходит в узле, к которому присоединяются пропуски. Но достаточно ли, что пропуски попадают в верный узел, для обоснования включения их в анализ? Ранее уже было отмечено, что в ходе экспериментов фиксировались также и изменения, происходящие в структуре дерева. Наиболее важное наблюдение, полученное в ходе исследования, состоит в следующем: изменения, условно названные «порчей» дерева, встречаются как при неверном, так и при верном определении пропусков.

Результаты экспериментов: последствия включения пропусков в анализ

При проведении экспериментов и сравнении полученных деревьев с исходными были выделены четыре вида возможных изменений в структуре дерева, которые происходят при включении пропусков в анализ (или четыре случая порчи дерева).

1. *Изменение структуры узлов, находящихся на той же глубине, что и узел с пропусками* (далее – *параллельные изменения*). Под такими изменениями подразумевается объединение в единый узел тех категорий предиктора, которые в исходном дереве располагались в разных узлах и имели статистически значимые различия в условных распределениях отклика, или же – напротив – разбиение на большее число узлов единого узла с несколькими категориями, у которых в исходном дереве статистически значимых различий в распределении отклика не наблюдалось.

2. *Появление узла, не заложенного в исходном дереве* (далее – *наличие* «*мусорного*» *узла*). Наличие этого изменения свидетельствует о том, что была найдена статистически значимая связь там, где она не предполагалась изначально; это изменение – следствие статистической ошибки 1-го рода.

В случае экспериментов с предиктором, располагающимся рядом с корнем, такие узлы могли появляться на более низкой глубине, а в случае предиктора, находящегося вдали от корня – только на той же глубине, т.е. параллельно узлам этой переменной. Этот факт накладывает ограничения на сравнение случаев предикторов на разной глубине: вероятность появления мусорных узлов при анализе предиктора определяется тем, сколько предикторов находится «под ней», т.е. глубже. Мы вывели следующее эмпирическое правило: предикторы, находящиеся на максимально возможной глубине, могут иметь мусорные узлы только на той же глубине; предикторы, находящиеся не на первом и не на последнем месте по глубине, могут иметь мусорные узлы как на той же глубине,

так и ниже; предикторы рядом с корнем – только ниже. Анализ и сравнение далее строится с учетом этого ограничения: при дальнейшем описании различий по этому виду порчи дерева имеются в виду именно ситуации, когда переменная находится «на первом месте» (рядом с корнем, на глубине 1) и «на последнем месте» (на максимально возможной глубине).

1. *Отсутствие заложенного в исходном дереве узла* (далее – *отсутствие узла*). Это вид порчи дерева «обратен» предыдущему и иллюстрирует ситуацию, когда статистически значимая связь не была найдена там, где она была заложена. Это изменение – следствие статистической ошибки 2-го рода. На этот вид порчи дерева также распространяется ограничение, описанное выше: для предиктора у корня вероятность отсутствия какого-либо узла будет выше, поскольку эта вероятность связана с узлами «ниже» этого предиктора, а для предиктора на максимально возможной глубине соответствующая вероятность будет ниже, поскольку узлы могут отсутствовать только на той же глубине.

2. *Изменение глубины расположения предикторов* (далее – *изменение глубины*). Этот вид порчи дерева иллюстрирует ситуацию, когда полученное после включения пропусков в анализ дерево вообще не воспроизводит заложенную структуру связей: предикторы меняются местами, а большинство получаемых узлов не соответствует ни одному из сочетаний предикторов, имеющихся в исходном дереве. При этом сами пропущенные значения располагаются не в одном узле, а рассредоточиваются в разных узлах по всему дереву – поскольку предикторы изменяют свое положение, а разбиение на узлы происходит совершенно иначе, чем в исходном дереве.

Частота встречаемости всех видов порчи дерева в целом, выраженная в доле экспериментов, представлена в *табл*. 3. Наиболее часто в проведенных экспериментах встречались мусорные узлы, так как этот вид порчи дерева характерен для предикторов на разной глубине, а наименее часто среди видов порчи дерева встречалось изменение глубины, поскольку оно более свойственно для случаев,

Таблица 3

ДОЛЯ ЭКСПЕРИМЕНТОВ С РАЗЛИЧНЫМИ ВИДАМИ ПОРЧИ ДЕРЕВА

Вид порчи дерева	Доля экспериментов
Параллельные изменения	0,468
Наличие мусорных узлов	0,545
Отсутствие узла	0,305
Изменение глубины	0,186

когда пропуски внедряются в предиктор рядом с корнем (хотя для переменной не у корня такое изменение дерева тоже возможно).

Так или иначе испорченными оказалась большая часть полученных деревьев – 75%. Для того чтобы проводить сравнение последствий включения пропусков в анализ в экспериментах с разными условиями, была введена мера оценки степени порчи дерева – индекс порчи дерева (ИПД), который рассчитывался следующим образом:

$$h = \sum_{i=1}^{4} w_i * m_i,$$

где h – ИПД в абсолютных значениях, m – наличие или отсутствие определенного вида порчи дерева, w – вес вида порчи дерева, который рассчитывался как $(1 – p)$, где p – оценка вероятности появления того или иного вида порчи дерева (доля экспериментов, где присутствовал этот вид порчи дерева). Таким образом, наибольший вес имели те виды порчи дерева, которые встречались реже всего.

Для того чтобы получить ИПД в относительном выражении, его значение в абсолютном выражении было поделено на максимально возможное (т.е. на сумму коэффициентов). Относительный ИПД показывает, на сколько «процентов» испорчено то или иное дерево. Максимально испорченными считались деревья, где на-

блюдалось изменение глубины расположения предикторов, так как этот вид порчи дерева сопровождается всеми остальными, а конечное дерево при этом нисколько не соответствует исходному. Средние значения относительного ИПД и доверительные интервалы для них представлены в *табл*. 4.

Результаты показывают, что наиболее благоприятная ситуация наблюдается в случае экспериментов с деревом, точность которого составляет 100%, а пропуски внедряются в переменную, расположенную вдали от корня: деревья при таких экспериментах не портились вовсе. Но следует понимать, что с подобной высокоточной моделью дерева пропуски объединяются в отдельный узел, уменьшая таким образом точность предсказаний пропорционально доле пропущенных значений. Однако важно, что все оставшиеся узлы полностью воспроизводят связи, заложенные в данных.

Для модели дерева с точностью 75%, когда пропуски также внедряются в предиктор не у корня, результаты ухудшаются. Как и в случае с определением пропусков по узлам, наиболее испорченные деревья встречаются в случае высокой доли пропусков, а для низкой и средней доли пропусков статистически значимой разницы в значениях индекса нет. Если предиктор находится на максимально возможной глубине (т.е. в дереве она «последняя»), то пропуски портят дерево незначительно, поскольку воздействие пропусков распространяется лишь на параллельные узлы. Иными словами, для такого предиктора порча дерева происходит лишь на глубине этой же самой переменной. Тем не менее и параллельные изменения, и наличие мусорных узлов, и отсутствие заложенных узлов – виды порчи дерева, характерные для случая предиктора на максимально возможной глубине – могут приводить к неверным выводам, а узлы, находящиеся на этой глубине, могут быть получены как следствие статистических ошибок.

Независимо от доли пропущенных значений, деревья портятся серьезнее, если пропуски находятся в предикторе рядом с корнем. Это обусловлено описанными выше причинами: поскольку именно

Таблица 4

СРЕДНИЕ ЗНАЧЕНИЯ ОТНОСИТЕЛЬНОГО ИПД

Доля пропусков	Положение предиктора											
	не у корня (глубина 3)						у корня (глубина 1)					
	Точность исходного дерева, %											
	75			100			75			100		
	Средн.	Н. гр. ДИ	В. гр. ДИ	Средн.	Н. гр. ДИ	В. гр. ДИ	Средн.	Н. гр. ДИ	В. гр. ДИ	Средн.	Н. гр. ДИ	В. гр. ДИ
10	0,178	0,153	0,203	0,0	0,0	0,0	0,269	0,234	0,304	0,102	0,067	0,137
25	0,206	0,173	0,239	0,0	0,0	0,0	0,440	0,391	0,489	0,128	0,095	0,161
50	0,287	0,246	0,328	0,0	0,0	0,0	1	1	1	1	1	1

из этого предиктора «исходят» все другие предикторы в дереве, он наиболее сильно подвержен всем изменениям, и в том числе самому «весомому» из них – изменению глубины расположения переменных. Причем в случае, когда доля пропусков очень высока (50%), предикторы всегда меняются местами, даже если исходная точность дерева достигает 100%. Таким образом, если исследователь решает включить в анализ переменную с такой высокой долей пропущенных значений, велика вероятность, что пропуски уменьшат реальную силу связи между предиктором и откликом, а переменная в итоге окажется «не на своем месте». Проблема здесь заключается также и в том, что исследователь, не имея информации о реальной, исходной структуре связей в данных, из полученной модели дерева может делать и содержательно неверные выводы – о значимых сочетаниях предикторов или о силе связи между переменными, и, следовательно, может получать артефактные результаты.

В случаях с низкой и средней долей пропусков ситуация улучшается, однако индекс порчи дерева все равно находится на более высоком уровне, чем для предиктора, который расположен на максимально возможной глубине. Исследователю необходимо понимать, что если пропуски находятся рядом с корнем или посередине дерева, т.е. в любом месте дерева, но не на максимальной глубине, то включение их в анализ могло повлечь изменения по всему дереву (на более глубоких ответвлениях). Интенсивность этих изменений зависит как от исходной точности дерева, так и от доли пропущенных значений.

Вернемся к вопросу: достаточно ли только попадания пропусков в верный узел, чтобы можно было «доверять» решению CHAID? Ответ на этот вопрос зависит от целей проводимого анализа. Если исследователь нацелен только на прогноз (классификацию или регрессию), то CHAID успешно справляется с этой задачей и при наличии пропусков: общий процент правильных предсказаний модели в каждой серии экспериментов был близок к исходному – независимо от того, верно или неверно определялись пропуски в узлы.

Однако если речь идет о поиске содержательно важных сочетаний предикторов, анализе эффектов взаимодействия и построении верных содержательных[1] выводов, то в этом случае следует иметь в виду, что чаще всего дерево, полученное при наличии в данных пропусков, испорчено – независимо от того, в верном узле оказываются пропуски или нет. Так, для экспериментов, которые завершились верным определением пропусков, среднее значение ИПД составило 0,24 (0,22–0,26), а для экспериментов с неверным определением пропусков – 0,63 (0,58–0,68). В реальных исследовательских ситуациях наибольшая сложность при использовании модели, полученной с включением пропусков, заключается в том, что исследователю неизвестно, какова исходная структура связей, верно ли определены пропуски и насколько испорченным получилось дерево. Усугубляется эта проблема и тем, что у исследователя нет каких-либо индикаторов того, что дерево испорчено: так, например, внутри серий проводимых экспериментов не наблюдалось статистически значимых различий между испорченными и не испорченными деревьями ни в общем проценте правильных предсказаний конечной модели – он всегда «стремится» к соответствующему значению в исходном дереве и незначительно варьируется в зависимости от доли пропусков (*табл.* 5), ни в устойчивости[2] моделей (*табл.* 6).

[1] Содержательный вывод в настоящем контексте противопоставляется статистическому выводу, который получен в результате проверки статистической гипотезы. Подразумевается, что содержательный вывод может быть неверным при верном статистическом выводе, если в ходе проверки имела место статистическая ошибка, наличие которой при использовании пропусков в CHAID было доказано в ходе описываемого эксперимента.

[2] Устойчивость модели измерялась как $(1 - |d|)$, где d – разница в показателе Risk между конечным деревом и деревьями, полученными при процедуре кросс-валидации. Risk – величина, показывающая общую долю ошибочных предсказаний. Таким образом, если устойчивость равна единице, это означает, что модель верно предсказывает одинаковую долю наблюдений при переносе ее на разные подвыборки. Чем ближе значение к единице – тем устойчивее дерево.

Таблица 5

СРЕДНИЕ ЗНАЧЕНИЯ ОБЩЕГО ПРОЦЕНТА ПРАВИЛЬНЫХ ПРЕДСКАЗАНИЙ

		Положение предиктора											
		не у корня (глубина 3)						у корня (глубина 1)					
		Точность исходного дерева, %											
		75			100			75			100		
Доля пропусков		Среднее	Н. гр. ДИ	В. гр. ДИ	Среднее	Н. гр. ДИ	В. гр. ДИ	Среднее	Н. гр. ДИ	В. гр. ДИ	Среднее	Н. гр. ДИ	В. гр. ДИ
10	Не испорчено	0,745	0,744	0,746	0,991	0,991	0,991	0,737	0,736	0,738	0,980	0,978	0,982
	Испорчено	0,745	0,745	0,745	–	–	–	0,738	0,737	0,739	0,979	0,978	0,980
25	Не испорчено	0,742	0,741	0,743	0,977	0,976	0,978	0,725	0,722	0,728	0,946	0,944	0,948
	Испорчено	0,743	0,742	0,744	–	–	–	0,723	0,722	0,724	0,947	0,945	0,949
50	Не испорчено	0,739	0,738	0,740	0,955	0,954	0,956	–	–	–	–	–	–
	Испорчено	0,738	0,736	0,740	–	–	–	0,698	0,697	0,699	0,895	0,893	0,897

Таблица 6

СРЕДНИЕ ЗНАЧЕНИЯ ПОКАЗАТЕЛЯ УСТОЙЧИВОСТИ КОНЕЧНЫХ МОДЕЛЕЙ

Доля пропусков		Положение предиктора											
		не у корня (глубина 3)				у корня (глубина 1)							
		Точность исходного дерева, %											
		75			100			75			100		
		Среднее	Н. гр. ДИ	В. гр. ДИ	Среднее	Н. гр. ДИ	В. гр. ДИ	Среднее	Н. гр. ДИ	В. гр. ДИ	Среднее	Н. гр. ДИ	В. гр. ДИ
10	Не испорчено	0,993	0,991	0,995	0,999	0,999	0,999	1	1	1	1	1	1
	Испорчено	0,993	0,992	0,994	–	–	–	1	1	1	1	1	1
25	Не испорчено	0,990	0,988	0,992	0,999	0,998	1	0,999	0,998	1	1	1	1
	Испорчено	0,992	0,991	0,993	–	–	–	1	1	1	1	1	1
50	Не испорчено	0,990	0,988	0,992	1	1	1	–	–	–	–	–	–
	Испорчено	0,992	0,991	0,993	–	–	–	0,999	0,999	0,999	1	1	1

Ни высокая точность полученного дерева, ни его устойчивость не могут быть основанием отрицать возможную порчу дерева.

Заключение

При решении проблемы наличия пропусков в данных выбор социолога обычно ограничен двумя альтернативами: исключить их, сократив доступную для анализа выборку, или искусственно заполнить, сохранив исходный объем наблюдений. Обе альтернативы обладают своими недостатками, повышающими актуальность и ценность использования таких методов анализа, которые способны работать с пропущенными значениями «как есть». К таким методам относятся многие алгоритмы деревьев решений. Так, при использовании наиболее распространенного в социологии алгоритма деревьев решений – CHAID – не происходит замены пропущенных значений на валидные: вместо этого пропуски рассматриваются как единая категория отклика, которая при построении модели присоединяется к наиболее похожему по распределению отклика узлу. И хотя возможность работы с пропущенными значениями в литературе определяется как уникальное преимущество деревьев решений, до сих пор в публикациях отсутствовали доказательства корректности включения пропусков в модели «напрямую», без заполнения.

В настоящем исследовании предпринята попытка с помощью статистического эксперимента установить, насколько корректно CHAID определяет пропущенные значения по узлам дерева, и выяснить, к каким последствиям приводит включение пропусков в модель. Анализ строился в разрезе трех условий экспериментов: расположения предиктора по отношению к корню, точности исходного дерева и доли пропусков.

Результаты экспериментов показали, что в целом CHAID корректно определяет пропуски по узлам: в большинстве случаев узлы, к которым присоединяются пропуски, располагаются на

корректной глубине, объединяют верные категории предиктора и содержат правильное распределение отклика, т.е. воспроизводят структуру связей, заложенную в данных. Наихудшие результаты в присоединении пропусков к категориям наблюдаются в случае, когда их доля очень велика (50%) – независимо от прочих условий.

Тем не менее в большинстве случаев включение пропусков в анализ сопровождается изменениями в структуре дерева – различными видами его порчи. Было выделено четыре возможных вида порчи дерева: структурные изменения в узлах на той же глубине, что и узел с пропусками, наличие мусорных узлов, отсутствие закладываемых узлов и изменение глубины расположения предикторов. Наиболее серьезно порче подвержены деревья в случаях, когда пропуски находятся в предикторах, располагающихся не на максимально возможной глубине, а, например, у корня или посередине дерева. Включение в анализ пропусков по таким предикторам может повлечь за собой изменения в структуре всех узлов, располагающихся ниже. Если же предиктор находится на максимально возможной глубине (т.е. в дереве он «последний»), то пропуски портят дерево незначительно, поскольку воздействие пропусков распространяется лишь на параллельные узлы.

В большинстве ситуаций анализ методом CHAID выполняет эксплораторные функции, и исследователю неизвестно, какова исходная структура связей, верно ли определены пропуски и насколько испорченным получилось дерево. При этом у исследователя нет каких-либо индикаторов, доказывающих, что дерево испорчено: было показано, что ни высокая точность, ни устойчивость модели не могут быть основанием отрицать возможную порчу дерева.

При принятии решений, стоит ли включать пропуски в анализ методом CHAID, следует учитывать следующие факторы.

1. *Цель анализа*. Если исследователь нацелен только на прогноз (классификацию или регрессию), то CHAID успешно справляется с этой задачей и при наличии пропусков. Если же речь идет о поиске содержательно важных сочетаний предикторов, анализе

эффектов взаимодействия и построении верных содержательных выводов, то в этом случае высока вероятность получить результаты, оборачивающиеся артефактами метода.

2. *Положение переменной с пропусками в структуре дерева*. Если пропуски находятся не «внизу» дерева, т.е. максимально глубоко, а в любом другом месте, то порча дерева может быть очень серьезной – вплоть до полного несоответствия реальной структуре связей. Чем ближе переменная к корню – тем интенсивнее портится дерево.

3. *Доля пропущенных значений*. Решение, полученное при слишком высокой доле пропусков (50%), чаще, чем в половине случаев не соответствует заложенной структуре связей в данных. Вероятность получить при таком условии адекватное решение очень низка.

Подобные результаты свидетельствуют о том, что возможность работы с пропущенными значениями в деревьях решений напрасно определяется в литературе как преимущество этих методов: риск получения неверных, ложных, ошибочных выводов существует – так же как и при искусственном заполнении пропусков. Необходимо отметить, что дизайн эксперимента включал в себя простую структуру дерева с тремя предикторами и дихотомическим откликом, а пропуски в экспериментах обладали полностью случайным характером. Следует предположить, что усложнение структуры дерева (какими и бывают модели в реальных исследованиях) и нарушение требований к случайному характеру пропусков лишь усугубят последствия включения их в анализ.

ЛИТЕРАТУРА

1. *Rubin D.B.* Inference and Missing Data // Biometrika. 1976. Vol. 63. P. 581–592.

2. *Ratner B.* Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. Boca Raton: CRC Press, 2012.

3. *Doove L.L., van Buuren S., Dusseldorp E.* Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects // Computational Statistics and Data Analysis. 2014. Vol. 72. P. 92–104.

4. *Von Hipple P.T.* How to Impute Interactions, Squares, and Other Transformed Variables // Sociological Methodology. 2009. Vol. 39. No. 1. P. 265–291.

5. *Dempster A.P., Rubin D.B.* Incomplete Data in Sample Surveys. Vol. 2: Theory and Annotated Bibliography. New York: Academic Press, 1983.

6. *Allison P.D.* Missing Data. Thousand Oaks, CA: Sage, 2002.

7. *Rokach L., Maimon O.* Decision Trees // Data Mining and Knowledge Discovery Handbook. Boston: Springer, 2010. P. 165–192.

8. *Kenett R., Salini S.* Modern Analysis of Customer Surveys: with Applications using *R*. Chichester: Wiley, 2012.

9. *Kass G.V.* An Exploratory Technique for Investigating Large Quantities of Categorical Data // Applied Statistics. 1980. Vol. 29. No. 2. P. 119–127.

10. *Quinlan J.R.* Unknown Attribute Values in Induction. Proceedings of the Sixth International Machine Learning Workshop. New York: Morgan Kaufmann Publishers Inc., 1989. P. 164–168.

11. *Gentle J.E., Härdle W.K., Mori Y.* Handbook of Computational Statistics: Concepts and Methods. Berlin: Springer, 2012.

12. *Gesser-Edelsburg A., Zemach M., Lotan T., Elias W., Grimberg E.* Perceptions, Intentions and Behavioral Norms that Affect Pre-license Driving among Arab Youth in Israel // Accident Analysis & Prevention. 2018. Vol. 111. P. 1–11.

13. *Ritschard G.* CHAID and Earlier Supervised Tree Methods. Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences / Ed. J. McArdle, G. Ritschard. London: Routledge. 2013. P. 48–74.

14. *Breiman L.* Random Forest. Machine Learning. 2001. Vol. 45. P. 5–32.

Приложение 1

Процедура генерирования базы данных для проведения экспериментов

Генерирование базы данных производилось в Microsoft Excel. В начале этой процедуры вероятностным образом (с помощью команды СЛУЧМЕЖДУ) создавались исходные три предиктора. Затем каждый из них был преобразован в набор фиктивных переменных, соответствующих категориям, которые были заложены в узлах. Так, для номинального предиктора n1 были созданы три фиктивные переменные, обозначающие категории «1 или 3», «2 или 5», «4», для предиктора *n*2 – фиктивные переменные «1 или 2», «3 или 4», а интервальный предиктор *i*1 был перекодирован в три фиктивные переменные, соответствующих значениям «меньше или равно 5», «от 5 до 10 включительно», «больше 10». Следующим этапом при гене-

рировании базы было создание переменных взаимодействия, отвечающих за крайние узлы, которые и формируют дерево. Для этого созданные ранее фиктивные переменные перемножались в соответствии с сочетаниями в этих узлах: например, для получения узла 7 были перемножены переменные «4» по исходной *n*1, «меньше или равно 5» по *i*1, «1 или 2» по *n*2; для узла 6 были перемножены переменные «4» по *n*1 и «больше 10» по *i*1 и т.д. Полученные переменные также принимали значения 0 и 1. Затем каждой из полученных переменных взаимодействия присваивался положительный или отрицательный коэффициент – в зависимости от того, единичные или нулевые значения содержатся в соответствующем крайнем узле (это определяется самим исследователем). При этом значение коэффициента, в отличие от знака, не играет существенной роли. Это объясняется математически: даже самое маленькое положительное значение коэффициента затем превратит экспоненту в число, хоть сколько-нибудь больше единицы (например, если коэффициент равен 0,0001, то *exp* = 1,0001). Тогда при расчете вероятностей значений отклика (а потом и самих его значений) числитель для положительного коэффициента всегда будет составлять больше половины знаменателя и вероятность будет больше 0,5 или – для отрицательного коэффициента – всегда будет составлять меньше половины знаменателя и вероятность – меньше 0,5. После этого значения переменных взаимодействия умножались на коэффициент и суммировались – таким образом было получено значение логита для каждого из сочетаний признаков, т.е. для каждого из наблюдений. Затем осуществлялся переход от исходной формы записи уравнения логистической регрессии к экспоненциальной через вычисление экспоненты в степени значения логита. Наконец, с помощью этого значения экспоненты и формулы отношения вероятностей стало возможно рассчитать вероятности наступления единичного события для каждого из наблюдений и затем присвоить им одно из двух значений отклика. Полная схема генерирования базы данных представлена на *рис*. 6.

Затем в дереве использовались исходные значения предикторов; фиктивные переменные и переменные взаимодействия присутствовали только на этапе генерирования базы. При этом переменная *n*1 целенаправленно была сделана неравномерной, 60% наблюдений в ней приходилось на значение «4» – для того чтобы соответствующий «переходный» узел был достаточно наполнен.

Приложение 2

Пример фрагмента syntax для проведения экспериментов с точностью исходного дерева 75%, долей пропусков 10%, внедряемых в переменную на глубине 3.

```
COMPUTE rand_1=RV.BERNOULLI(0.1).
EXECUTE.
COMPUTE n2_miss_1=n2.
EXECUTE.

DO IF (rand_1 = 1).
RECODE n2_miss_1 (ELSE=SYSMIS).
END IF.
EXECUTE.

* Decision Tree.
TREE y_75 [n] BY n1 [n] i1 [s] n2_miss_1 [n]
 /TREE DISPLAY=TOPDOWN NODES=STATISTICS
BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO
 /DEPCATEGORIES USEVALUES=[0 1]
 /PRINT MODELSUMMARY CLASSIFICATION RISK TREETABLE
/SAVE NODEID
 /METHOD TYPE=CHAID
 /GROWTHLIMIT MAXDEPTH=3 MINPARENTSIZE=20
MINCHILDSIZE=20
 /VALIDATION TYPE=CROSSVALIDATION(10)
OUTPUT=BOTHSAMPLES
 /CHAID ALPHASPLIT=0.05 ALPHAMERGE=0.05
SPLITMERGED=NO CHISQUARE=PEARSON CONVERGE=0.001
 MAXITERATIONS=1000 INTERVALS=3
 /COSTS EQUAL
 /MISSING NOMINALMISSING=MISSING.
USE ALL.
COMPUTE filter_$=(rand_1 = 1 & NodeID_1 = 9).
VARIABLE LABELS filter_$ ‘rand_1 = 1 (FILTER)’.
VALUE LABELS filter_$ 0 ‘Not Selected’ 1 ‘Selected’.
FORMATS filter_$ (f1.0).
```

```
FILTER BY filter_$.
EXECUTE.
FREQUENCIES VARIABLES=n2
 /STATISTICS=MODE
 /ORDER=ANALYSIS.
FREQUENCIES VARIABLES=y_75
 /STATISTICS=MODE
 /ORDER=ANALYSIS.
FILTER OFF.
USE ALL.
<…>
```

Zhuchkova Svetlana,
National Research University Higher School of Economics (NRU HSE), Moscow, lana_lob@mail.ru

Rotmistrov Alexey,
National Research University Higher School of Economics (NRU HSE), Moscow, alexey.n.rotmistrov@gmail.com

Handling missing data with CHAID: results of a statistical experiment
The paper addresses an approach to working with a missing data «as is», implying that missing data can be viewed as a separate variable category. This approach is different from alternative approaches – deleting observations with missing data or replacing missing data with valid information. One of the methods that work with missing data «as is» is CHAID. CHAID refers to the decision trees class of methods. This method is relevant for researchers dealing with categorical variables and nonlinear associations. We did not find an answer to the question what are the advantages and limitations of the CHAID approach comparing to the mentioned alternatives from previous research, although tree models with missing data are often found in empirical studies. To start a discussion considering this issue, we conducted a series of statistical experiments on generated data organized into three predictors of categorical and interval measure type. The empirical finding was that this method correctly distributes missing data in tree's nodes, but in most cases the inclusion of missing data into analysis is accompanied by changes in tree's structure, and therefore there is a risk of obtaining incorrect, false, erroneous conclusions. The paper also provides recommendations on what factors should be considered when deciding whether to include missing data in an analysis «as is».
Keywords: categorical variables, CHAID, classification tree, decision tree, interaction effects, missing data, statistical experiment.

References
1. Rubin D.B. Inference and missing data, *Biometrika*, 1976, 63, 581–592.
2. Ratner B. *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data*. Boca Raton: CRC Press, 2012.
3. Doove L.L., van Buuren S., Dusseldorp E. Recursive partitioning

for missing data imputation in the presence of interaction effects, *Computational Statistics and Data Analysis*, 2014, 72, 92–104.
4. Von Hipple P.T. How to impute interactions, squares, and other transformed variables, *Sociological Methodology*, 2009, 39(1), 265–291.
5. Dempster A.P., Rubin D.B. *Incomplete Data in Sample Surveys*. Vol. 2: Theory and Annotated Bibliography. New York: Academic Press, 1983.
6. Allison P.D. *Missing Data.* Thousand Oaks, CA: Sage, 2002.
7. Rokach L., Maimon O. *Decision Trees. Data Mining and Knowledge Discovery Handbook.* Boston: Springer, 2010. P. 165–192.
8. Kenett R., Salini S. *Modern Analysis of Customer Surveys: with Applications using R*. Chichester: Wiley, 2012.
9. Kass G. V. An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, 1980, 29(2), 119–127.
10. Quinlan J.R. *Unknown Attribute Values in Induction. Proceedings of the Sixth International Machine Learning Workshop*. New York: Morgan Kaufmann Publishers Inc., 1989. P. 164–168.
11. Gentle J.E., Härdle W.K., Mori Y. *Handbook of Computational Statistics: Concepts and Methods*. Berlin: Springer, 2012.
12. Gesser-Edelsburg A., Zemach M., Lotan T., Elias W., Grimberg E. Perceptions, intentions and behavioral norms that affect pre-license driving among Arab youth in Israel, *Accident Analysis & Prevention*, 2018, 111, 1–11.
13. Ritschard G. "CHAID and Earlier Supervised Tree Methods", in: McArdle J., Ritschard G. (ed.) *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*. London: Routledge, 2013. P. 48–74.
14. Breiman L. Random forest, *Machine Learning*, 2001, 45, 5–32.