К.А. Тенишева, С.С. Савельева, Д.А. Александров
(*Санкт-Петербург*)

ПРИМЕНЕНИЕ МЕТОДА УСЛОВНЫХ ДЕРЕВЬЕВ РЕШЕНИЙ К МОДЕЛИРОВАНИЮ ВЫБОРА РОДИТЕЛЯМИ ШКОЛЫ

Представлен новый для социологии подход к изучению выбора – применение метода условных деревьев решений. Подробно разбирается логика метода и его преимущества на примере анализа выбора родителями школы в двух районах Санкт-Петербурга. Показывается, что деревья решений хорошо подходят для выделения групп, следующих разным стратегиям принятия решений. Метод может быть эффективным инструментом моделирования и интерпретации логики принятия решений. Он выигрывает в сравнении с традиционным моделированием при помощи логистической регрессии, поскольку позволяет оценить гомогенность предпочтений (выборов) полученных групп, а не просто выделять ключевые для выбора факторы. Предлагается в научных и прикладных социальных исследованиях, посвященных изучению сложного выбора, сочетать регрессионный анализ с методом деревьев решений.

Ключевые слова: логистическая регрессия, классификация, моделирование выбора, метод условных деревьев решений, выбор школы.

Ксения Алексеевна Тенишева – кандидат социологических наук, младший научный сотрудник Научно-учебной лаборатории «Социология образования и науки», Национальный исследовательский университет «Высшая школа экономики» в Санкт-Петербурге. E-mail: ktenisheva@hse.ru.
Светлана Сергеевна Савельева – зам. заведующего Научно-учебной лаборатории «Социология образования и науки», Национальный исследовательский университет «Высшая школа экономики» в Санкт-Петербурге. E-mail: ssavelieva@hse.ru.
Даниил Александрович Александров – кандидат биологических наук, профессор, заведующий Научно-учебной лаборатории «Социология образования и науки», Национальный исследовательский университет «Высшая школа экономики» в Санкт-Петербурге. E-mail: dalexandrov@hse.ru.

Введение

Одна из ключевых тематик в социологии образования – изучение выбора. Исследователи определяют предикторы выбора образовательных учреждений – от детского сада до вуза или колледжа, выбора ступени образования, способа получения конкретного уровня образования.

Традиционно моделируют именно бинарные выборы. При этом даже сложные ситуации принятия решения, когда актор рассматривает несколько доступных альтернатив, сводятся к последовательности попарных – бинарных – сравнений. Отчасти это связано с методологией, которая с давних пор навязывает именно такую «бинарную» оптику для изучения выбора, отчасти – с теорией и общими представлениями о том, как устроен выбор для акторов. К примеру, в теории рационального выбора процесс принятия решения представляется в том числе как последовательное попарное сравнение альтернатив для их ранжирования.

В социальных науках выбор анализируется с помощью регрессий, в основном применяются бинарная и мультиномиальная логистическая регрессия – для ситуаций с двумя и с множеством возможных исходов соответственно.

Мы предлагаем новый подход к моделированию образовательных выборов – деревья решений. В сфере классификации этот метод не нов, однако в социологии его практически не применяли до сих пор. Социальные исследователи традиционно используют логику деревьев решений, не прибегая при этом к специальным современным техникам, призванным рассчитывать как ключевые узлы, так и вероятностные характеристики возможных исходов, а продолжая прибегать к регрессионному анализу. Ярким примером такого подхода может послужить направление изучения образовательного выбора, продолжающее работы Р. Брина и Дж. Голдторпа [1].

Вместе с тем давно разработанный метод деревьев решений наилучшим образом подходит для моделирования таких ситуаций.

На данный момент он позволяет осуществлять классификацию как в ситуации бинарных выборов (из двух опций), так и в случаях выбора из большего количества альтернатив.

Здесь мы представляем подробный обзор метода деревьев решений: его принципов, алгоритма и логики. Затем мы представляем анализ нашего кейса – выбор родителями типа школы в двух районах города – с помощью данного метода. Наконец, на тех же данных мы строим классическую логистическую регрессию для содержательного сопоставления результатов применения методов разного типа.

Методы анализа данных в исследованиях неравенства и выбора образования

Выбор школы – та область исследований неравенства в образовании, где используется весь арсенал методов исследований. Некоторые исследования основаны на анализе исключительно качественных данных, ярким их примером служат исследования выбора школы в Великобритании [2]. Однако большинство работ основано на анализе больших массивов количественных данных [3; 4]. В 1980–1990-х гг. ученые зафиксировали, что неравенство сохраняется, несмотря на возникшую экспансию начального и общего образования [5]. Появились первые попытки объяснить этот феномен на количественном уровне.

В рамках одного направления образовательные достижения рассматриваются в качестве решений (да или нет) в отношении получения образования следующего уровня [6; 7]. Такой способ изучения связи социального происхождения и достигнутого уровня образования стал популярным благодаря Р. Мэа и получил название «изучение переходов между образовательными уровнями» (educational transitions tradition) [6]. Под переходом понимается поступление на образовательную ступень следующего уровня/класса после окончания предыдущей, например, когда ученики из одного класса переходят в следующий класс или с одного уровня на другой (из школы

в училище/университет), или же вовсе перестают учиться [8]. Исследователи этой традиции стремятся обнаружить моменты, когда социальное происхождение более всего влияет на образовательное решение. Поскольку анализ представляет собой изучение факторов, влияющих на бинарный выбор (например, отправляется учиться в колледж выпускник школы или нет), самый оптимальный метод для такого исследования – построение бинарных логистических регрессий. Модель Мэа была долгое время самым популярным способом и до сих пор применяется в изучении образовательной стратификации.

Однако эта модель часто критикуема за упрощение реальности, ведь образовательная траектория зачастую организована нелинейно и непоследовательно [9]. Образовательные системы имеют несколько качественно разных программ внутри одного уровня, а также обеспечивают возможность достичь определенного уровня образования разными путями. Это заставляет исследователей обращаться к мультиномиальной логистической регрессии, где зависимая переменная может выражать определенный набор выборов [9]. Также в зависимости от конкретных задач исследования применяется линейная регрессия.

Регрессионное моделирование остается самым распространенным на данный момент способом анализа неравенства в образовании в целом и выборе школы, в частности, хотя в последнее время в этой области появляется все больше новых техник анализа. В частности, в США проводятся исследования с экспериментальным дизайном, где иногда численность выборки доходит до полумиллиона [10]. Мы предлагаем применить новый метод – деревья решений – и демонстрируем, какие он имеет преимущества.

Метод деревьев решений

Деревья решений были созданы для решения задачи классификации. Эта задача – одна из наиболее типичных для многих наук, включая медицину, криминологию и социологию. Медики

могут заранее классифицировать, например, подойдет ли почка реципиенту в ситуациях с высоким риском отторжения [11], криминологи – по психологическому профилю определят склонность индивида к совершению правонарушений [12], социологи – принадлежность индивидов к тем или иным социальным группам. Деревья решений выступают достаточно простыми классификаторами и отлично подходят для описанных выше задач. На основе имеющихся данных они предсказывают категории – тип болезни, социальную группу, класс, тип школы, т.е. одну из доступных категорий выбора. Деревья решений, разумеется, не единственный существующий на данный момент классификатор. Однако он выгодно отличается от других методов по двум критериям.

1. Деревья решений прозрачнее, чем «черные ящики» сложносоставных классификаторов, основанных на нейронных сетях[1]. Этот классификатор дает возможность увидеть логику, по которой алгоритм присваивает классы.

2. Этот метод лучше справляется с классификацией неполных данных, а также намного менее требователен к исходным данным (нет требований к нормальности, независимости предикторов и т.п.). Деревья решений показывают лучшие результаты классификации на данных с пропусками по сравнению с дискриминантным анализом [13; 14]. Благодаря своей нечувствительности к типу распределения переменных и отсутствию требования независимости предикторов, деревья оказываются более удобным и надежным методом классификации в случаях большого количество (возможно) взаимосвязанных предикторов, по сравнению с линейной или логистической регрессией – в частности не страдают от проблем мультиколлинеарности и гетероскедастичности[2].

[1] Имеются в виду такие классификаторы, как байесовские нейронные сети. О том, почему они представляют из себя черный ящик и какими методами его можно приоткрыть, см. [15].

[2] О сопоставлении данных алгоритмов см., к примеру: [16].

В классической методике деревья решений обучаются – «тренируются» на данных, для которых известны и характеристики наблюдений, и значения зависимой переменной, какую мы предсказываем. Иными словами, мы располагаем полными данными, позволяющими оценить точность классификации и шансы ошибок для разных категорий моделируемой переменной. Затем уже обученное дерево легко применяется для предсказания категорий зависимой переменной на данных, где имеются все предикторы, но неизвестны значения зависимой переменной.

Первые алгоритмы деревьев решений появились в 1960-х гг. (первой считается работа Моргана и Сонкуиста, изданная в 1963 г. [17]), тогда были разработаны два базовых алгоритма для деревьев классификации (classification trees)[1] – CART и C4.5, представленные одновременно Брейманом с коллегами [18] и Кинланом [19;20]. Подробнее с историей деревьев решений можно ознакомиться в статье Штробл и др. [21]. В конце 90-х и 2000-х гг. метод приобрел большую популярность из-за резкого роста интереса к майнингу данных, поскольку одной из его базовых задач является классификация, и возросших компьютерных мощностей, позволяющих обрабатывать огромные массивы данных в сжатые сроки. Разработанные ранее алгоритмы (AID, CHAID, FACT, QUEST, C4.5, CART) активно совершенствовались. Кроме того, были созданы новые альтернативные алгоритмы, включая CTree, реализованный в пакете party и значительно доработанный для пакета partykit[2] (эти алгоритмы реализованы в программной среде R).

Логика работы любого дерева решений очень проста: классифицировать данные, последовательно «задавая вопросы» о

[1] В статье «деревья решений» и «деревья классификации» употребляются как синонимы. Формально деревья классификации – один из типов более широкого семейства деревьев решений, которое также включает регрессионные деревья предсказаний (regression/prediction trees), появившиеся несколько позже.

[2] Описание пакета и его применения см.: https://cran.r-project.org/web/packages/partykit/vignettes/partykit.pdf (дата обращения: 21.11.2018).

характеристиках наблюдений, связанных с исследуемым феноменом [22]. В зависимости от этих характеристик наблюдения последовательно делятся на подгруппы.

Группа разнородных наблюдений представляются в виде узла (node) и в соответствие с вопросом (характеристикой) делится на дочерние узлы (child nodes) – по одному на каждый вариант ответа. В самом элементарном представлении для каждого вопроса существует два варианта ответа: да и нет. Каждый из них ведет к своему узлу. Этот узел «задает» следующий вопрос, а ответы на него либо снова порождают дочерние узлы, либо приводят к финальному узлу, который также называют листом (leaf). Этот узел представляет категорию, к которой относится выделенная группа наблюдений. Дерево строится от корня (root) – первого узла, от которого начинается разделение – к листьям, т.е. финальным категориям. Финальные узлы содержат информацию об определенной деревом категории: условную вероятность принадлежности к категории, относительную частоту наблюдений данной категории в выделенной подгруппе, вероятность ошибки классификации. Наиболее современные пакеты, включая partykit, позволяют выводить информацию разного формата по финальным узлам.

На *рис.* 1 представлено дерево для категоризации условий погоды для игры на свежем воздухе. Всего учитывается три параметра: прогноз, который предсказывает погоду трех типов: солнечная, пасмурная или дождливая; влажность, измеряемая по интервальной шкале; наличие ветра, бинарный фактор. Предсказывается переменная игра, для которой возможны всего два значения: да или нет.

Характеристикой, стоящей у корня и наилучшим образом делящей выборку на подгруппы, является прогноз погоды. Он принимает три значения и таким образом порождает три ветки. Две из них (солнечно и дождливо) ведут к дочерним узлам, одна (пасмурно) сразу приводит к финальному узлу и предсказывает категорию «да» для игры. Если прогноз обещает солнечную погоду, дочерним узлом становится влажность, интервальная переменная,

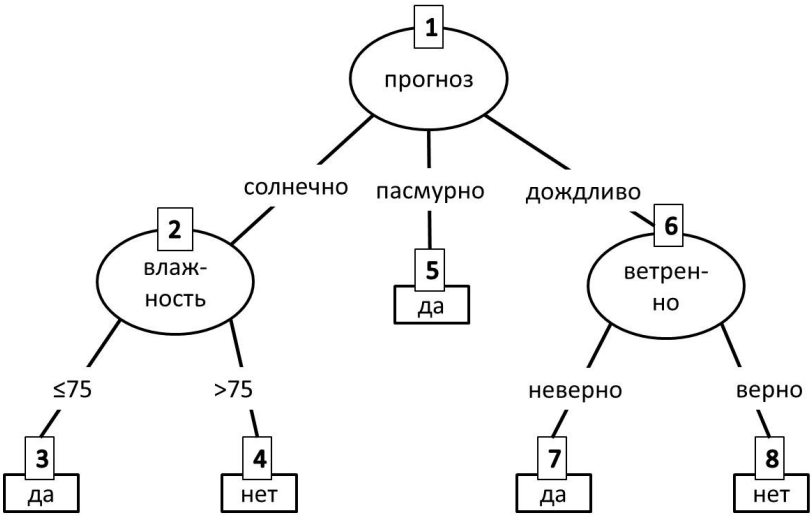

Рис. 1. Базовый пример дерева решений для пакета partykit

Источник: [20].

для которой путем перебора выделяется пороговое значение. Если влажность будет не больше 75, финальным узлом становится «да». Если влажность будет выше 75, мы получаем категорию «нет». Если же корневая категория принимает значения «дождливо», дочерним узлом становится наличие ветра. Если ветер есть (верно), мы получаем «нет» в качестве финального узла; если ветра нет (неверно), финальная категория для игры – «да».

Как видно из данного примера, современные алгоритмы деревьев решений работают с переменными любого типа – и в качестве предикторов, и в качестве предсказываемого фактора. Это значит, что в процессе категоризации разделение наблюдений на подгруппы может происходить и по бинарным показателям (таким как пол), и по мультиномиальным (этничность), и по категориальным/порядковым (уровень образования), и по интервальным (доход). По умолчанию дочерних узлов будет два, но дерево можно вручную

модифицировать, вменив ему разделение на 3 и более дочерних узла (именно поэтому в дереве решений, представленном на *рис*. 1, после корня идет разделение на три категории, а не на две).

Для предикторов всех типов, кроме бинарных, будет определено простое логическое правило разделения на подгруппы – больше/меньше(/равно) для переменных с упорядоченными значениями и «относится к группе значений А»/«относится к группе значений Б» для номинальных переменных. Подробнее принцип выбора значения для разделения на подгруппы для переменных разных типов будет показан далее в разделе «Алгоритмы». Также деревья решений предсказывают любой тип зависимой переменной. Достаточно очевидно, что они хорошо справляются с предсказанием не только бинарных, но и мультиномиальных и категориальных переменных. Также разработаны алгоритмы построения регрессионных деревьев для случаев, когда зависимая переменная является интервальной, например, CART и C4.5 [23].

В целом, до начала 2000-х гг. алгоритмы бинарного разделения страдали от трех серьезных общих проблем.

1. Предпочтение переменных с большим количеством значений и, как следствие, большим количеством вариантов разбивки. Интервальная переменная всегда выбирается для узла в сравнении с бинарной, даже в ущерб объяснительной силе модели: до недавнего времени в процессе разделения не было никаких показателей статистической значимости.

2. Сильная склонность к переобучению: не имея четкого критерия, определяющего количество конечных групп, этот алгоритм дает избыточное число групп, что вредит интерпретируемости и ценности результатов.

3. Эти алгоритмы очень чувствительны к структуре тренировочных данных. Иногда даже небольшие изменения переменных могут привести к серьезным изменениям в логике построения дерева решений.

Опубликованный в 2006 г. алгоритм условного рекурсивного разделения (conditional partitioning) для деревьев решений [24; 25;

26], представленный сначала в пакете party и затем доработанный в пакете partykit программной среды R, избавился от этих проблем и представляет собой один из наиболее надежных для задач социальных наук на данный момент методов построения деревьев решений. Поэтому в дальнейшем в тексте речь идет именно об условном рекурсивном разделении из упомянутых выше пакетов. Мы не приводим сравнение существующих алгоритмов и типов деревьев решений, а лишь демонстрируем логику применения наиболее современного и полезного (на наш взгляд) для социологии образования подхода.

Далее мы разбираем, какие алгоритмы лежат в основе отбора переменных для узлов, определения значения переменной, по которому идет разделение на дочерние узлы, и выбора критерия остановки дерева решений (описание основано на [27]).

Алгоритмы

Алгоритм состоит из трех простых шагов.

1. Выбор переменной, по которой будет проходить разделение на дочерние узлы, среди включенных в анализ предикторов на основе значения α, одновременно определяющего и порог значимости при отборе ковариаты (стандартно 0,05), и размер итогового дерева. Алгоритм останавливается, если подтверждается глобальная нулевая гипотеза об отсутствии связей, т. е. ни одна из включенных в модель переменных значимо не взаимосвязана с зависимой.

2. Выбор порогового значения переменной для разделения наблюдений на подгруппы (для бинарной переменной это всегда будет 1 vs. 0; для других типов переменных задействуется алгоритм перебора возможных вариантов разделения, предпочитающий порог, дающий наиболее статистически различные группы).

3. Рекурсивное повторение первых двух шагов до остановки алгоритма.

Разделение шагов 1 и 2, т.е. выбора переменной и значения для разделения на подгруппы, позволяет избежать проблем с предпо-

чтением предикторов с наибольшим количеством категорий или самым большим числом пропущенных значений. Этот алгоритм также дает наиболее интерпретируемую структуру дерева решений.

Разберем подробно каждый из шагов.

Отбор переменной для узла

На первом шаге необходимо выбрать переменную, по которой будет проводиться разделение всей выборки на две подгруппы. Для этого надо решить проблему независимости, т.е. определить, какие из включенных в модель предикторов значимо взаимосвязаны с зависимой переменной и какая из этих взаимосвязей предпочтительна для создания узла. Глобальная нулевая гипотеза заключается в предположении об отсутствии значимой взаимосвязи зависимой переменной хотя бы с одной из ковариат (включенных в модель предикторов). Если мы не можем отвергнуть нулевую гипотезу на заданном пороге значимости α, рекурсия – и, следовательно, построение дерева – останавливается. Если мы можем отвергнуть нулевую гипотезу, мы измеряем взаимосвязь между зависимой переменной Y и каждым из предикторов $X_j; j \in [1; ...; m]$. Таким образом производится проверка частных нулевых гипотез – об отсутствии связи Y с каждым из предикторов.

Проверка частных гипотез основана на тестах перестановок (об их применении к проблеме классификации см.: [28]). В содержательном плане эта проверка демонстрирует случайность либо неслучайность взаимосвязи зависимой переменной и предиктора. Если тест показывает, что такое соотношение значений зависимой переменной и предиктора маловероятно, т.е. неслучайно, частная нулевая гипотеза отвергается.

Полученные для разных предикторов тестовые статистики нельзя сопоставлять напрямую, если все ковариаты не измеряются по одной шкале. Наиболее простым способом сопоставления результатов тестов для включенных в анализ переменных становится

сравнение их значений *p*, чей размер не зависит от типа и дробности шкалы, по которой измеряется переменная. В итоге критерием выбора переменной, по которой будет проходить разделение на дочерние узлы, служит наименьшее значение критерия значимости *p*.

Итак, на первом шаге глобальная нулевая гипотеза отвергается в том случае, когда минимальное значение *p* из набора *p*, полученных для ковариат, оказывается меньше назначенного порога α, в противном случае алгоритм останавливается. Переменная с минимальным уровнем значимости, согласно частным гипотезам, становится узлом, по которому проходит разделение на подгруппы. Таким образом, параметр α можно считать определяющим размер конечного дерева: чем меньше значение α, тем меньшее количество предикторов будет ему удовлетворять.

Выбор значения для разделения на подгруппы

Следующим шагом для алгоритма становится выбор конкретного значения переменной, отобранной на шаге 1, по которому будет проходить разделение на две подгруппы (для создания двух дочерних узлов). Для этого также применяются тесты перестановок. Адекватность выбранного значения для разделения на подгруппы (goodness of split) оценивается через двухвыборочные тесты, которые представляют собой частный случай статистических тестов, применяемых для проверки взаимосвязи зависимой переменной и ковариаты на первом шаге (отборе переменной). В основе поиска значения для разделения лежит перебор. Иначе говоря, для всех значений, которые может принимать ковариата, последовательно проводится сравнение: выборка каждый раз делится на две подгруппы по данному значению ковариаты, для них вычисляется значение тестовой статистики, затем выборка делится на две подвыборки по следующему значению предиктора. Для итогового разделения наблюдений на две подгруппы (и создания дочерних узлов) выбирается значение ковариаты, дающее

наибольшее значение тестовой статистики. Выбирается разделение, которое дает не просто значимое различие между двумя подгруппами, но и максимальное различие (discrepancy) тестовой статистики между ними, что позволяет совершать выбор между несколькими порогами, дающими значимый результат.

Избежать появления патологических узлов, т.е. узлов, содержащих слишком мало и к тому же не эффективных для категоризации наблюдений, можно просто установить порог для минимального количества наблюдений или суммы весов кейсов в каждой из сопоставляемых подвыборок, обращаясь к функции ctree_control(). Например, следующая команда устанавливает минимальную сумму весов в каждой из двух подгрупп равной 20:

```
> ctree_control(minsplit = 20)
```

Пропущенные значения

Важно понимать, как алгоритм поступает с пропущенными значениями. Пропуски в зависимой переменной не допускаются, их необходимо удалить (или, наоборот, заполнить) до начала анализа. На данный момент в пакете partykit нет встроенных алгоритмов заполнения пропусков, которые автоматически работали бы для функции ctree, создающей дерево решений. Если пропуски содержатся в предикторе, применяется техника создания суррогатных узлов. Идея этого метода заключается в поиске других признаков в базе данных, которые могли бы указать на принадлежность к одной из двух образуемых подгрупп в ситуации, когда главный признак – значение, по которому производится разделение – отсутствует. Суррогатные узлы также создаются для тех случаев, когда значение предиктора отсутствует не в тренировочных, а в тестовых данных (подробнее о методе см.: [29])

Суррогатные переменные служат достаточно эффективным методом для деревьев, особенно если в переменной меньше 10% пропусков. Если пропусков больше, лучший результат показывает

заполнение пропусков методом множественной импутации[1]. Но различия в успешности этих методов невелики, так что ими можно пренебречь. Главное выгодное отличие импутации заключается в том, что она позволяет восстанавливать пропуски в зависимой переменной, и это можно делать заранее, до начала моделирования с помощью деревьев решений [31].

Далее покажем, как можно применить метод условных деревьев решений к анализу выбора родителями школы на оригинальных эмпирических данных.

Исследование выбора школы: описание дизайна и эмпирических данных

Сбор и анализ данных проводился Научно-учебной лабораторией «Социология образования и науки» (НИУ ВШЭ в Санкт-Петербурге) в рамках проектов, поддержанных Программой фундаментальных исследований НИУ ВШЭ и Российским гуманитарным научным фондом[2]. Для изучения выбора школы было отобрано два района Петербурга, отличающихся по своим основным характеристикам, – Василеостровский и Невский. Сбор данных проводился в период с 2013 по 2016 г. Целью было опросить не менее половины школ каждого из районов, они выбирались из списков всех школ районов случайным образом. В Невском районе были опрошены все выбранные школы, в Василеостровском получено два отказа от гимназий с наиболее высоким рейтингом – они не попали в выборку. В итоге в Василеостровском районе был опрошен 581 человек в 21-й школе (из 30-ти школ в районе), а в Невском – 474 родителя в 13-ти школах (из 19-ти школ левобережной части района). Опрос проводился студентами и сотрудниками НИУ ВШЭ, которые

[1] О множественной импутации см., например: [30].

[2] Проект № 16-03-00802 «Дифференциация школ и образовательный выбор: школа и родители» 2016–2018 гг.

встречали родителей у школы, проводили экспресс-интервью и записывали ответы в анкету. Опросный инструмент содержал как закрытые, так и открытые вопросы, которые были посвящены деталям выбора школы: например, сколько вариантов рассматривали, как долго и как именно выбирали, как сравнивали школы, что привлекало, на какие источники информации опирались в своем выборе. Затем уточнялись намерения менять школу после окончания 4 класса (вопрос задавался всем родителям, так как многие планируют перевести ребенка в другое учебное заведение после окончания начальной школы уже на момент поступления) и планы относительно дальнейшего образования детей, нацеленность семей на получение высшего образования. Отдельный блок вопросов был посвящен социально-демографическим характеристикам семьи: фиксировался уровень образования родителей и их социально-экономический статус.

Основные переменные, вошедшие в итоговые модели:

– расстояние до школы в количестве минут, которые тратит на дорогу ребенок, интервальная переменная принимает значения от 1 до 60 (timetoget);

– уверенность в том, что сегодняшний школьник в дальнейшем получит высшее образование, категориальная переменная: принимает значение 1 – «обязательно получит», 2 – «возможно, получит», 3 – «вряд ли получит высшее образование» (togetHE);

– собираются ли переводить ребенка в другую школы после окончания начальной, категориальная переменная: принимает значения 1 – «да, обязательно», 2 – «возможно, переведем», 3 – «вряд ли», 4 – «нет, не собираемся переводить» (tracking);

– уровень образования матери, категориальная переменная: 1 – среднее, 2 – начальное профессиональное, 3 – среднее специальное, 4 – высшее образование (mothedu);

– социально-экономический статус представлен в виде международного социально-экономического индекса ISEI'08, основанного на международном классификаторе профессий (ISCO-08).

В соответствии с ним существующие профессии ранжируются в зависимости от их престижа в обществе и требуемого уровня образования. В этой системе классификации самые высокие баллы, более 80-ти, присваиваются высококвалифицированным профессионалам, а самые низкие, менее 20 – необразованным «синим воротничкам». Таким образом, это интервальная переменная, принимающая в данной выборке значения от 17 до 89-ти (ISEI_mother);

– район, категориальная переменная, отражающая район опроса: 1 – Василеостровский район, 2 – Невский район (district).

Переменные, отражающие значимые характеристики школ для родительского выбора, принимали значение 1, если отмечались в качестве важной характеристики, и значение 0 – в противном случае. В модели вошли следующие переменные:

– охрана школы, безопасность учеников (import_secur);

– статус школы (гимназия, лицей, специализированная школа) (import_status);

– высокие школьные результаты ЕГЭ (import_ege);

– наличие кружков и секций в школе (import_curic);

– этнический состав в школе (import_ethnic);

– культурный уровень одноклассников ребенка (import_cult).

Основные описательные статистики по переменным представлены в *табл*. 4 и 5 в *Приложении*.

Нашей целью является моделирование выбора типа школы. Нас интересует, какие характеристики школы и семьи связаны с выбором обычной общеобразовательной школы либо школы повышенного статуса. Под школой повышенного статуса подразумеваются гимназии, лицеи, а также школы с углубленным изучением предметов (в Василеостровском районе таких школ 57%, в Невском, где в целом намного меньшее разнообразие школьной системы, – 29%). Зависимой переменной становится переменная «статус школы», и она не имеет пропущенных значений. Предикторы содержат пропуски: наибольшее их количество в перемен-

ной «социально-профессиональный статус матери» (310 кейсов). Однако для алгоритма деревьев решений это не проблема. Как было описано выше, если переменная с пропусками становится узлом, для нее создается набор суррогатных узлов, которые помогут провести классификацию[1].

Применение метода деревьев решений к анализу выбора школы

Дерево решений для предсказания выбора типа школы было построено с помощью функции ctree() из пакета partykit. Она создает условное дерево решений и возвращает объект класса partynode, содержащий список всех объектов, необходимых для описания результатов: корня, узлов, порогов, финальных категорий и суррогатных узлов.

Дерево решений задается формулой стандартного для генерализованных регрессионных моделей вида, в качестве аргументов функции ctree() мы указываем зависимую переменную (schtype2) и предикторы; они разделяются знаком тильда (~). Затем в аргументе data указываются данные, на которых строится модель (sch). С помощью аргумента ctree_control() можно вводить дополнительные параметры расчета дерева: ограничивать ею размер, количество финальных узлов, устанавливать границу по количеству наблюде-

[1] Мы не можем быть абсолютно уверенными в том, что нашем случае пропуски значений случайные, а не систематические. Это может привести к неточным результатам классификации. Однако отдельные исследования показывают, что восстановление значений оказывается эффективным даже если исследователь принял неслучайные пропуски (MNAR) за случайные (MAR) [32], что дает основания полагаться на алгоритм суррогатных узлов в нашем случае.

С точки зрения валидности и точности, на данный момент наиболее надежным методом работы с пропусками в моделях с деревьями решений, по всей видимости, оказались ансамбли (сочетание двух и более методов восстановления пропусков). Например, сочетание метода MICE и суррогатных узлов, заложенных в алгоритм пакета ctree [33]. Мы намерены уделить большее внимание точности обработки пропущенных значений в нашей дальнейшей работе.

ний и весам в дочерних и финальных узлах, задавать количество суррогатных узлов и проч. В данном случае мы ограничили максимальное количество суррогатных узлов двумя: maxsurrogate = 2.

```
> cfitNVc<-ctree(schtype2~timetoget + togetHE + tracking+mothedu
+ import_secur + import_status + import_ege + import_curic +
import_ethnic + import_cult + district + ISEI_mother, data=sch,
control=ctree_control(maxsurrogate = 2))
```

Структуру полученного дерева решений можно отобразить двумя способами: списком узлов и листьев либо в виде графа. Список дает всю необходимую исследователю информацию: порядок разделения на узлы, границы значений, по которым проводится разделение, количество случаев в каждом из финальных узлов и доля ошибочных классификаций. Граф же оказывается более наглядным для презентации, при этом на нем также можно отобразить всю необходимую информацию. Ниже мы подробно описываем результаты моделирования, апеллируя как к представлению дерева списком, так и к графическому объекту. Мы рекомендуем на этапе анализа обращаться к обоим способам вывода дерева решений.

Структуру дерева решений списком можно изучить, просто введя название полученного объекта.

```
> cfitNVc
Model formula:
schtype2 ~ timetoget + togetHE + tracking + mothedu + import_secur
+ import_status + import_ege + import_curic + import_ethnic +
import_cult + district + ISEI_mother
Fitted party:
[ 1 ] root
| [ 2 ] status in Неважно
| | [ 3 ] district in Василеостровский
| | | [ 4 ] ege in Неважно
| | | | [ 5 ] mothedu in Среднее, Начальное профессиональное,
```

Среднее специальное: сош (n = 144, err = 27.1%)

| | | [6] mothedu in Высшее: сош (n = 211, err = 49.8%)

| | | [7] ege in Важно: повыш статус (n = 39, err = 23.1%)

| | [8] district in Невский

| | | [9] timetoget <= 25: сош (n = 302, err = 8.3%)

| | | [10] timetoget > 25: сош (n = 14, err = 35.7%)

| [11] status in Важно

| | [12] district in Василеостровский

| | | [13] togetHE in Да, обязательно: повыш статус (n = 170, err = 14.1%)

| | | [14] togetHE in Возможно, Вряд ли: повыш статус (n = 58, err = 44.8%)

| | [15] district in Невский

| | | [16] timetoget <= 11: сош (n = 102, err = 39.2%)

| | | [17] timetoget > 11: повыш статус (n = 56, err = 32.1%) Number of inner nodes: 8

Number of terminal nodes: 9

Мы видим перечисление всех выделенных алгоритмом узлов в той же последовательности, в какой они возникают в дереве решений – от корня к финальным узлам (листьям). Для каждого узла указывается значение переменной, по которому выборка разделилась на две подгруппы. Цифра перед названием узла является его уникальным номером; единица всегда присваивается корню. Для финальных узлов (СОШ либо повыш. статус) указывается число кейсов, попавших в категорию, и процент ошибочной классификации, т. е. доля попавших в этот лист случаев, которые относятся к другой категории. Чем меньше доля ошибок, тем аккуратнее наша классификация и выше гомогенность предпочтений в выделенной подгруппе.

Графический вывод дерева решений проще для чтения и интерпретации, чем список. Полезными будут два типа вывода структуры дерева. Первый вариант дает представление об услов-

ной вероятности каждой из категорий зависимой переменной в каждом из финальных узлов, он отображен на *рис*. 2.

```
>plot((cfitNVc))
```

Второй вариант отображает вместо условных вероятностей в финальных узлах классификацию: количество наблюдений и вероятность ошибки. Для его построения используется аргумент as.simpleparty() – *рис*. 3.

```
>plot(as.simpleparty(cfitNVc))
```

В овалах отображаются предикторы, по которым происходит разделение. Каждый овал пронумерован и этот номер соответствует номеру узла в описанном выше выводе. Под названием узла указывается условие разделения. Также для каждого разделения указывается значимость теста на наличие связи зависимой переменной и предиктора (значение *p*). Представление нижнего ряда дерева – листьев – различается для двух типов графов.

В первом варианте (*рис*. 2) листья показаны в форме столбцов, отображающих условную вероятность каждого из двух исходов. Темно-серым закрашена область вероятности выбора школы повышенного статуса, светло-серым – вероятность выбора СОШ. Над каждым из листьев подписано количество кейсов, попавших в данную подгруппу.

Во втором варианте (*рис*. 3) в нижнем ряду дерева решений представлен результат классификации зависимой переменной. В каждом финальном узле (листе) подписан один из двух возможных исходов: СОШ или школа повышенного статуса, а также указана доля ошибок классификации. Чем ниже процент, тем выше точность определения, к какой категории принадлежит родитель, на основе выбранных предикторов.

В нашем случае корнем выбрана переменная «важен ли был статус школы» (status)[1]. Если родители ответили, что статус был

[1] Может показаться, что переменная, важен ли статус школы родителям, дублирует предсказываемую переменную – реально выбранный статус. Но это не так:

не важен, мы переходим по левой ветке ко второму узлу – району. Продолжая двигаться по левой ветке, мы получаем узел для родителей из Василеостровского района. Для этой подгруппы следующим фактором становится важность для родителей результатов ЕГЭ школы. Если ЕГЭ был не важен (левая ветка), мы переходим к последнему узлу: наличию у матери высшего образования. Если у матери нет высшего образования (начальное, среднее профессиональное), мы переходим к финальному узлу номер 5 и попадаем в категорию СОШ (144 случая) с вероятностью ошибки классификации 27,1%. Если же у матери есть высшее образование, мы получаем узел 6 (211 случаев), также СОШ, но со значительно большей долей неверной классификации: 49,8%.

Иначе говоря, последовательность «не важен статус школы» – Василеостровский район – «не важны результаты ЕГЭ» в любом случае приводят к категории СОШ. Но для матерей без высшего образования мы получаем гомогенную по своим предпочтениям группу с низким уровнем ошибки классификации. А для матерей, получивших высшее образование, группа оказывается гетерогенной, и для них определение выбранного статуса школы сравнимо с угадыванием: в половине случаев это будет СОШ, в половине – школа повышенного статуса.

Возвращаясь на шаг назад, к узлу номер 3 «важность ЕГЭ» мы получаем еще один финальный узел для тех, кто считает ЕГЭ важным: они классифицируются как выбирающие школы повышенного статуса (39 кейсов), с низким уровнем ошибки в 23,1%. Или иначе, семьи, которым не важен статус школы, из Василеостровского района, но считающие важным ЕГЭ школы, скорее отдадут ребенка в школу повышенного статуса – и это предсказание достаточно аккуратно.

внутри обеих групп (тех, кто считает статус важным и не важным) существует большая вариация по итоговому выбору школы. К тому же важность статуса не всегда значит важность повышенного статуса – некоторые родители ценят именно статус «стандартной» школы как менее строгой, требовательной к ребенку.

Мы также построили дерево решений, исключив переменную «важность статуса», и получили схожую структуру.

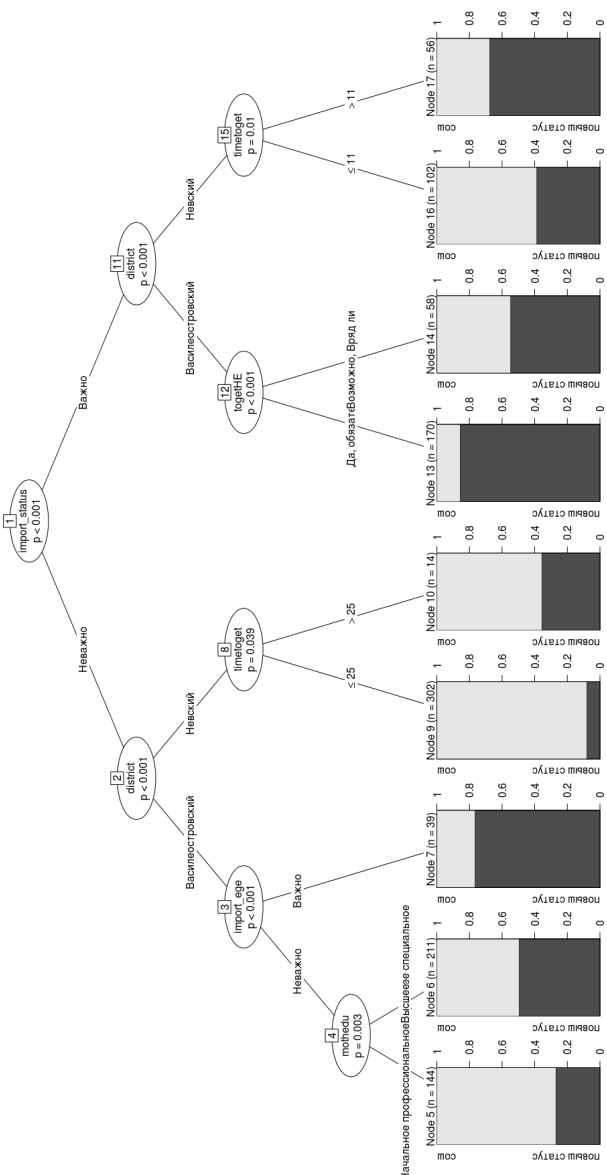

Рис. 2. Дерево решений с условными вероятностями выбора школы определенного типа

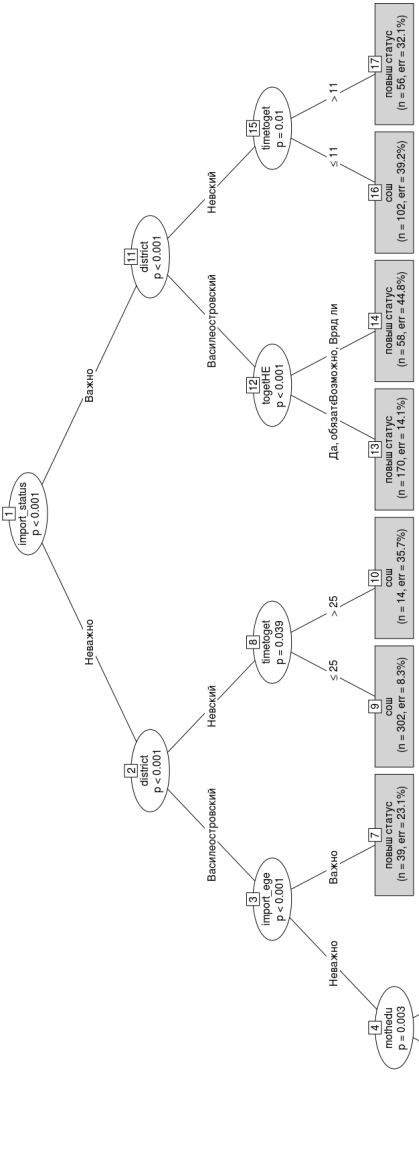

Рис. 3. Дерево решений с категоризацией выбора школы определенного типа

Если сделать еще шаг назад и вернуться к узлу 2 (район) и пойти по правой ветке, созданной для Невского района, мы получим следующий узел: время, затрачиваемое на дорогу до школы (timetoget), ведущее к двум листьям. Если дорога до школы занимает не больше 25 минут, выбор классифицируется как СОШ (302 случая) с высоким уровнем точности – всего 8,3% ошибок. Если же на дорогу уходит больше 25 минут, выбор снова классифицируется как СОШ (14 кейсов), но со значительно возросшим уровнем ошибок – 35,7%. Для тех, кому не важен статус школы, в Невском районе выделяется две группы родителей. Те, кто водит в школу неподалеку – и это практически всегда оказывается обычная школа. И те, кто водит в более удаленную от дома школу, хотя для них шансы выбрать школу повышенного статуса все равно ниже, чем шансы выбрать СОШ. Это объясняется структурными условиями района: в Невском районе просто намного меньше школ повышенного статуса, чем в Василеостровском районе. Поэтому для абсолютного большинства семей выбор такой школы активен, они должны хотеть отдать ребенка не в СОШ.

Давайте теперь посмотрим на правую ветку дерева решений. Она начинается с выбора категории «важно» для вопроса о статусе школы. Вторым узлом здесь, как и в левой части дерева, становится переменная «район». Для Василеостровского района (левая ветка) следующим предиктором становится представление родителей о том, получит ли их ребенок высшее образование (togetHE). Если они считают, что «обязательно получит» (левая ветка), мы попадаем в финальный узел номер 13, категория «школа повышенного статуса» (170 кейсов), с низким уровнем ошибки 14,1%. Если же родители менее уверены, что ребенок пойдет в вуз, их выбор все равно категоризируется как школа повышенного статуса (58 кейсов), но с ошибкой в 44,8% случаев.

Иначе говоря, родители, которым важен статус школы, живущие в Василеостровском районе и нацеленные на получение ребенком высшего образование, представляют собой еще одну

гомогенную группу, отдающую детей в школы повышенного статуса. Те же, кто не имеет таких образовательных притязаний, принадлежат к группе с разнообразными предпочтениями: чуть больше половины этих родителей выбирают школы повышенного статуса, и лишь чуть меньше половины отдают детей в СОШ.

Наконец, последняя ветка выделяется для тех, кто считает статус школы важным и живет в Невском районе. Для них определяющим фактором снова становится время, которое семья тратит на дорогу до школы. Если на дорогу уходит не больше 11 минут, выбором родителей становится СОШ (102 кейса), с заметным уровнем ошибок категоризации (39,2%). Если же дорога занимает больше 11 минут, семья скорее всего отдала ребенка в школу повышенного статуса (56 случаев), с чуть более высокой точностью категоризации – 32,1% ошибок.

Для Невского района, независимо от отношения родителей к важности статуса школы, ключевым фактором для определения, в школу какого типа семья отдала ребенка, служит время, уходящее на дорогу до школы. Из-за бедного образовательного ландшафта этого района, выбор в пользу СОШ или школы повышенного статуса реально определяется готовностью семьи водить ребенка в достаточно удаленную школу. И, как мы видим, это решение даже не связано с образовательным или социально-экономическим статусом семьи – они не были выбраны в качестве предикторов для категоризации выбора родителей Невского района.

Василеостровский район является контрастным кейсом с большим разнообразием очень доступных в территориальном плане учебных заведений. Так как здесь не стоит так остро вопрос дальности пути в школу, важность приобретают другие факторы – представление о важности академических результатов школы и наличие у матери высшего образования, т.е. от культурного капитала семьи.

Не все переменные, включенные в модель, попали в дерево решений. Это означает, что все остальные факторы, такие как важность этнического состава школы или наличие внеклассных занятий, не способы улучшить классификацию: они не снижают

уровень энтропии (гетерогенности) в полученных подгруппах и, следовательно, бесполезны для дерева решений. Это не означает, что родители не ориентируются на данные показатели. Скорее, речь о том, что на эти не вошедшие в дерево факторы могут ориентироваться родители из уже полученных подгрупп. Иначе говоря, представления о важности таких характеристик школы довольно однородны внутри полученных «совокупностей» родителей, поэтому не могут уточнить нашу классификацию.

В этом ключевое отличие деревьев решений от регрессии: в итоговую модель входят не все переменные, которые на первом шаге – отборе корня дерева – были значимо связаны с зависимой переменной. Остаются в модели только те факторы, которые не теряют значимости для последовательно выделяемых подвыборок. Регрессия же по определению дает список переменных, значимых на всей совокупности.

Мы можем проверить, какие из переменных изначально были значимо связаны с зависимой. Для этого есть два способа. Чтобы увидеть коэффициенты, по которым происходил отбор переменных, нужна функция nodeapply(). Она выдает значения $\log(1 - p)$, поскольку они более стабильны и лучше подходят для сравнения. Однако они не очень удобны для интерпретации, поэтому мы покажем таблицу значений p без логарифма. Для этого применяется функция sctest() (structural change test). В *табл.* 1 получаем статистики и значимость предикторов для первого узла (корня).

```
>library("strucchange")
>sctest(cfitNVc, node = 1)
```

Мы видим, что, к примеру, переменная «важна внеклассная деятельность» была значима на этом первом шаге ($p = 0{,}04283852$), однако показатель значимости оказался ниже для переменной «важен статус школы», которая и стала первой в структуре дерева решений. А культурный уровень одноклассников так и не вошел в итоговую модель.

Таблица 1

ТЕСТОВАЯ СТАТИСТИКА И ЗНАЧИМОСТЬ КАЖДОГО ИЗ ПРЕДИКТОРОВ ДЛЯ ПЕРВОГО УЗЛА

Переменная	statistic	p.value
timetoget	9,41	0,03
togetHE	19,66	6,67e–4
tracking	5,92	0,77
mothedu	47,19	3,78e–09
import_secur	11,23	0,01
import_status	144,51	0,00
import_ege	8,211	0,04
import_curic	8,45	0,04
import_ethnic	13,68	<0,01
import_cult	3,67	0,49
district	126,12	0,00
ISEI_mother	22,55	2,45e–05

В *табл*. 2 показана пропорция каждой из категорий (СОШ и школа повышенного статуса) в каждом из финальных узлов дерева решений. Также мы видим количество кейсов в каждом из узлов. Очевидно, лучше классифицируются листья, по которым видно четкое превалирование доли одной категории над другой, как для узлов 5 (СОШ), 9 (СОШ), 13 (школа повышенного статуса). Те же узлы, где пропорции близки к равным, не могут четко описать паттерн выбора типа школы (как узлы 6, 14).

Качество полученного дерева решений можно оценить, разделив выборку на тренировочную и тестовую. На тренировочной выборке дерево обучается, после чего на тестовой мы проверяем долю «угадывания» категорий объясняемой переменной на основе имеющихся переменных. В нашем случае мы создали тренировочную выборку, случайным образом отобрав 75% кейсов. Соответственно, в тестовой совокупности осталось 25%. В *табл*. 3 мы

Таблица 2

ПРОПОРЦИЯ КАТЕГОРИЙ СОШ И ПОВЫШЕННЫЙ СТАТУС В КАЖДОМ ФИНАЛЬНОМ УЗЛЕ, в %

№ узла	*N*	СОШ	Повышенный статус
5	144	72,92	27,08
6	211	50,24	49,76
7	39	23,08	76,92
9	302	91,72	8,28
10	14	64,29	35,71
13	170	14,12	85,88
14	58	44,83	55,17
16	102	60,78	39,22
17	56	32,14	67,86

Таблица 3

КЛАССИФИКАЦИЯ ВЕРНЫХ И ОШИБОЧНЫХ КАТЕГОРИЗАЦИЙ ТИПА ШКОЛЫ НА ОСНОВЕ ТЕСТОВОЙ ВЫБОРКИ

Тестовая выборка		Обучающая выборка		Итого
		СОШ	повышенный статус	
СОШ	*N*	133	60	193
	%	0,68	0,31	
Повышенный статус	*N*	25	56	81
	%	0,31	0,69	
Итого		158	116	274

видим долю правильных и неверных категоризаций для двух типов школы – СОШ и школа повышенного статуса. Наше дерево верно определяет выбор в пользу СОШ в 68,9% случаев и выбор школы повышенного статуса в 69,1% случаев.

Общий уровень точности (accuracy), рассчитанный как отношение правильных категоризаций к общему числу кейсов, составляет 68,97%. Это означает, что дерево не угадывает категории

слепо (тогда аккуратность была бы равна примерно 50%), однако классификация не является и абсолютно точной. Для оценки результата следует учитывать, что мы работаем с социальными данными сложной природы и ограниченным набором предсказывающих факторов. Следует признать, что для 30% случаев существуют иные, не учтенные в нашем исследовании, параметры, определяющие их выбор в пользу общеобразовательной школы или школы повышенного статуса.

Применение логистической регрессии к анализу выбора школы

Для сравнения нового метода деревьев решений с традиционными методами, мы проводим анализ выбора типа школы с помощью логистической регрессии. Мы включили в нее тот же набор переменных, который включен в дерево решений. Применен метод автоматического выбора предикторов «backwards selection». Изначально была построена полная модель, включающая все возможные предикторы. Затем алгоритм последовательно исключил все статистически незначимые факторы. Результаты моделирования представлены в *табл*. 4.

Полученная в результате применения логистической регрессии модель в некоторых аспектах заметно отличается от результатов, выдаваемых деревом решений. Как и в дереве, никакой роли в объяснении выбора типа школы не играет социально-профессиональный статус матери, а высшее образование матери повышает шансы на выбор школы повышенного статуса. Схожий с продемонстрированный деревом эффект оказывают и такие факторы, как время, которое уходит на дорогу до школы (дальше водят в школы повышенного статуса); представление о том, что статус школы важен так же как и результаты ЕГЭ. Воспроизводится также базовое различие между районами города: семьи из Невского района в принципе имеют меньше шансов отдать ребенка в школу повышенного статуса.

Таблица 4

РЕЗУЛЬТАТЫ ЛОГИСТИЧЕСКОГО РЕГРЕССИОННОГО АНАЛИЗА (ЗАВИСИМАЯ ПЕРЕМЕННАЯ – СТАТУС ШКОЛЫ)

Переменная	Коэффициент	Доверительный интервал
Время до школы	0,022**	(0,004, 0,040)
Получит ВО *Возможно*	−0,435**	(−0,791, −0,079)
Получит ВО *Вряд ли*	−0,980**	(−1,685, −0,275)
Смените школу *Возможно*	1,393***	(0,660, 2,126)
Смените школу *Вряд ли*	1,110**	(0,376, 1,845)
Смените школу *Нет*	1,508***	(0,817, 2,200)
Образование матери *Начальное профессиональное*	−1,482	(−3,459, 0,495)
Образование матери *Среднее специальное*	−0,027	(−0,679, 0,625)
Образование матери *Высшее*	0,724**	(0,122, 1,325)
Статус школы *Важно*	1,831***	(1,503, 2,160)
ЕГЭ *Важно*	0,915***	(0,517, 1,314)
Внешкол.занятия *Важно*	−0,464**	(−0,814, −0,115)
Этнический статус *Важно*	−1,060***	(−1,638, −0,482)
Район *Невский*	−1,916***	(−2,266, −1,566)
Константа	−1,951***	(−2,860, −1,041)
Наблюдений	723	
AIC	751,690	
Pseudo Rsq (CS)	0,314	
Pseudo Rsq (N)	0,421	
D	0,329	

Примечание. $^{*}p < 0{,}1$; $^{**}p < 0{,}05$; $^{***}p < 0{,}01$

Однако значимыми оказываются и другие аттитюдные вопросы о важности различных характеристик школы, которые не стали частью дерева решений. Мы видим, что шансы выбрать школу повышенного статуса ниже у семей, считающих важным наличие внеклассных занятий в самой школе, а также этнический состав школы. Родители, выбирающие привилегированную школу для ребенка, имеют меньше шансов планировать дальнейший переход в другое учебное заведение. Планы на получение высшего образования имеют прямую связь с шансами на выбор школы повышенного статуса: чем с больше уверенностью родители говорят, что ребенок получит в итоге высшее образование, тем выше шансы, что они отдали ребенка в гимназию или лицей.

Разумеется, общая логика, выявленная деревом решений, сохраняется и в регрессионной модели, но все же интерпретация результатов различается: появляются факторы, которые не попали в дерево решений, поскольку оказались недостаточно эффективными для разделения выборки на подгруппы. В определенной мере большее количество факторов проливают дополнительный свет на причины выбора родителями школы определенного типа. Значимость переменных в логистической регрессии говорит, что их связь с типом школы неслучайна и они могут быть использованы для предсказания шансов родителей на выбор школы того или иного типа. Если дерево демонстрирует логику принятия решения (последовательность факторов), то регрессия выявляет факторы, которые в целом различают родителей, выбирающих обычные школы, и тех, кто выбирает школы повышенного статуса.

Заключение

Мы подробно описали особенности метода деревьев решений и продемонстрировали аналитические возможности, которые дает его применение к объяснению выбора типа школы.

Можно сказать, что результаты регрессии и деревьев решений дополняют друг друга. Регрессия может дать представление о

более широком спектре предикторов, значимых для исследуемой совокупности. Однако результаты логистической регрессии, как и любой другой, сложно интерпретировать с точки зрения логики выбора, это просто набор характеристик, так или иначе с этим выбором связанных.

В этом отношении техника деревьев решений оказывается интереснее. Она демонстрирует иерархию предикторов, которую для исследований процессов принятия решений можно, с одной стороны, интерпретировать как иерархическую последовательность самой логики выборов. С другой стороны, дерево решений может скрыть от нас какой-то из значимых факторов, потому что на одном из этапов он оказался чуть менее значим, чем другой. Но вместе с тем оно дает нам подробное представление о том, какие группы родителей выделяются в вопросе выбора школы. Иначе говоря, обладающие какими характеристиками родители, судя по всему, действуют в одной логике выбора.

Ошибки классификации в дереве решений сами по себе становятся дополнительным источником информации о том, как родители совершают выбор. В случаях с низкой долей ошибок мы можем считать группу гомогенной в своих характеристиках и устойчивой в предпочтениях. В тех же случаях, когда дерево решений не может снизить уровень ошибок, мы получаем представление о группе, которая осталась достаточно гетерогенной: несмотря на общность ключевых характеристик, члены группы совершают выбор по-разному. Вероятно, нам как исследователям стоит обратить особое внимание именно на эти подгруппы и постараться выявить дополнительные факторы, зачастую не очевидные, которые влияют на исход принятия решения о выборе типа школы. Метод регрессии, в отличие от дерева решений, не дает нам подобной информации.

В социальных исследованиях, как научных, так и прикладных, для изучения кейсов сложного выбора мы предлагаем использовать оба метода. Пример анализа, продемонстрированный в статье, помогает построить модель логики принятия решений в любой сфере.

ЛИТЕРАТУРА

1. *Breen R., Goldthorpe J.H.* Explaining Educational Differentials: Towards a Formal Rational Action Theory // Rationality and Society. 1997. Vol. 9(3). P. 275–305.

2. *Ball S.J.* Good School/Bad School: Paradox and Fabrication // British Journal of Sociology of Education. 1997. Vol. 18 (3). P. 317–336.

3. *Taylor C.* Hierarchies and Local Markets: the Geography of the Lived Market Place in Secondary Education Provision // Journal of Education Policy. 2001. Vol. 16(3). P. 197–214.

4. *Kristen C.* School Choice and Ethnic School Segregation: Primary School Selection in Germany: Waxmann Verlag, 2003.

5. *Shavit Y., Blossfeld H.P.* Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Social Inequality Series. Boulder: Westview Press, 1993.

6. *Mare R.D.* Change and Stability in Educational Stratification // American Sociological Review. 1981. Vol. 1. P. 72–87.

7. *Shavit Y., Blossfeld H.-P.* Persistent Inequality: Changing Educational Attainment in Thirteen Countries // Social Inequality Series. ERIC, 1993.

8. *Lucas S.R.* Effectively Maintained Inequality: Education Transitions, Track Mobility, and Social Background Effects // American Journal of Sociology. 2001. Vol. 106(6). P. 1642–1690.

9. *Breen R., Jonsson J.O.* Analyzing Educational Careers: A Multinomial Transition Model // American Sociological Review. 2000. Vol. 65(5). P. 754–772.

10. *Cullen J.B., Jacob B.A., Levitt S.D.* The Impact of School Choice on Student Outcomes: an Analysis of the Chicago Public Schools // Journal of Public Economics. 2005. Vol. 89 (5–6). P. 729–760.

11. *Shaikhina T., Lowe D., Daga S., Briggs D., Higgins R., Khovanovaa N.* Decision Tree and Random Forest Models for Outcome Prediction in Antibody Incompatible Kidney Transplantation // Biomedical Signal Processing and Control. 2017. https://doi.org/10.1016/j.bspc.2017.01.012.

12. *Masias V.H., Valle M.A., Amar J.J., Cervantes M., Brunal G., Crespo F.A.* Characterising the Personality of the Public Safety Offender and Non-offender using Decision Trees: The Case of Colombia // Journal of Investigative Psychology and Offender Profiling. 2016. Vol. 13(3). P. 198–219.

13. *Feldesman M.R.* Classification Trees as an Alternative to Linear Discriminant Analysis // American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists. 2002. Vol. 119. P. 257–275.

14. *Karels T.J., Bryant A.A., Hik D.S.* Comparison of Discriminant Function and Classification Tree Analyses for Age Classification of Marmots // Oikos. 2004. Vol. 105(3). P. 575–587.

15. *Guidotti R., Monreale A., Ruggieri S., Turini F.* A Survey of Methods for Explaining Black Box Models // ACM Computing Surveys (CSUR). 2018. Vol. 51(5). P. 93.

16. *Markou E.* 3 Machine Learning Algorithms You Need to Know. URL: https://dzone.com/articles/3-machine-learning-algorithms-you-need-to-know (date of access: 21.11.2018).

17. *Morgan J.N., Sonquist J.A.* Problems in the Analysis of Survey Data, and a Proposal // Journal of the American Statistical Association. 1963. Vol. 58(302). P. 415–434.

18. *Breiman L., Friedman J.H., Olshen R.A., Stone C.J.* Classification and Regression Trees. New York: Chapman and Hall, 1984.

19. *Quinlan J.R.* Induction of Decision Trees // Machine Learning. 1986. Vol. 1(1). P. 81–106.

20. *Quinlan J.R.* C4.5: Programms for Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 1993.

21. *Strobl C., Malley J., Tutz G.* An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests // Psychological Methods. 2009. Vol. 14(4). P. 323–348.

22. *Kingsford C., Salzberg S.L.* What are Decision Trees? // Nature Biotechnology. 2008. Vol. 26. P. 1011–1013.

23. *Song Y.Y., Ying L.U.* Decision Tree Methods: Applications for Classification and Prediction // Shanghai Archives of Psychiatry. 2015. Vol. 27(2). P. 130–135.

24. *Hothorn T., Hornik K., Zeileis A.* Unbiased Recursive Partitioning: A Conditional Inference Framework // Journal of Computational and Graphical Statistics. 2006. Vol. 15(3). P. 651–674.

25. *Zeileis A., Hothorn T.* partykit: A Toolkit for Recursive Partytioning, 2012. URL: https://cran.r-project.org/web/packages/partykit/vignettes/partykit.pdf (date of access: 21.11.2018).

26. *Hothorn T., Hornik K., Zeileis A.* ctree: Conditional Inference Trees // The Comprehensive R Archive Network, 2015. URL: https://rdrr.io/rforge/partykit/f/inst/doc/ctree.pdf (date of access: 21.11.2018).

27. *Hothorn T., Hornik K., Strobl C., Zeileis A.* Party: A Laboratory for Recursive Partitioning, 2010. URL: https://cran.r-project.org/web/packages/party/vignettes/party.pdf (date of access: 21.11.2018).

28. *Golland P., Liang F., Mukherjee S., Panchenko D.* Permutation Tests for Classification // International Conference on Computational Learning Theory. Springer, Berlin, Heidelberg, 2005. P. 501–515.

29. *Therneau T.M., Atkinson E.J.* An Introduction to Recursive Partitioning Using the RPART Routines, 2018. URL: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf (date of access: 21.11.2018).

30. *Фабриканта М.С.* Модель-ориентированный подход к отсутствующим значениям: множественная импутация в многоуровневой регрессии посредством R (на примере анализа опросных данных) // Социология: методология, методы, математическое моделирование. 2016. № 41. С. 7–29.

31. *Feelders A.* Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation? // European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg, 1999. P. 329–334.

32. *Janssen K.J., Donders A.R.T., Harrell Jr. F.E., Vergouwe Y., Chen Q., Grobbee D.E., Moons K.G.* Missing Covariate Data in Medical Research: to Impute is Better than to Ignore // Journal of Clinical Epidemiology. 2010. Vol. 63(7). P. 721–727.

33. *Valdiviezo H.C., Van Aelst S.* Tree-based Prediction on Incomplete Data Using Imputation or Surrogate Decisions // Information Sciences. 2015. No. 311. P. 163–181.

Приложение 1

Описательные статистики по переменным

Таблица 4

ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ ДЛЯ ИНТЕРВАЛЬНЫХ ПЕРЕМЕННЫХ

Переменная	*N*	Среднее	Стндартное отклонение	Мин	Макс
timetoget	1,087	10,940	8,667	0	60
ISEI_mother	786	50,628	14,730	17	89

Таблица 5

ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ ДЛЯ КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ

Переменная	Значения	*N*	%	Кумулятивный %
schtype2	сош	636	58,0	58,0
	повыш статус	460	42,0	100,0
	Всего	1096	100,0	
togetHE	Да, обязательно	731	69,2	69,2
	Возможно	263	24,9	94,1
	Вряд ли	62	5,9	100,0
	Нет, не получит	0	0,0	100,0
	Всего	1056	100,0	

Продолжение табл. 5

Переменная	Значения	*N*	%	Кумулятивный %
tracking	Да, обязательно	71	6,7	6,7
	Возможно	213	20,2	26,9
	Вряд ли	194	18,4	45,3
	Нет, не собираемся	578	54,7	100,0
	Всего	1056	100,0	
mothedu	Среднее	103	9,5	9,5
	Начальное профессиональное	10	0,9	10,4
	Среднее специальное	296	27,3	37,7
	Высшее	677	62,3	100,0
	Всего	1086	100,0	
import_secur	Не важно	864	78,8	78,8
	Важно	232	21,2	100,0
	Всего	1096	100,0	
import_status	Не важно	709	64,8	64,8
	Важно	386	35,2	100,0
	Всего	1095	100,0	
import_ege	Не важно	874	79,7	79,7
	Важно	222	20,3	100,0
	Всего	1096	100,0	
import_curic	Не важно	818	74,6	74,6
	Важно	278	25,4	100,0
	Всего	1096	100,0	
import_ethnic	Не важно	986	90,0	90,0
	Важно	110	10,0	100,0
	Всего	1096	100,0	

Окончание табл. 5

Переменная	Значения	*N*	%	Кумулятивный %
import_cult	Не важно	902	82,4	82,4
	Важно	193	17,6	100,0
	Всего	1095	100,0	
district	Василеостровский	622	56,8	56,8
	Невский	474	43,2	100,0
	Всего	1096	100,0	

Tenisheva Ksenia
Sociology of Education and Science Laboratory, National Research University Higher School of Economics (NRU HSE), Saint Petersburg, ktenisheva@hse.ru

Savelieva Svetlana
Sociology of Education and Science Laboratory, National Research University Higher School of Economics (NRU HSE), Saint Petersburg, ssavelieva@hse.ru

Alexandrov Daniel
Sociology of Education and Science Laboratory, National Research University Higher School of Economics (NRU HSE), Saint Petersburg, dalexandrov@hse.ru

Method of conditional inference decision trees for modeling parental school choice

In the article we suggest a method for analyzing situations of choice, which is novel for sociology – the method of conditional decision trees. We describe the logic of the method, applying it to the case of parental choice of school in two urban districts of Saint-Petersburg. We show that decision trees work well for detecting groups that follow different decision making strategies. This can be an efficient tool for modeling and interpretation of the logic of choice. The method outperforms the traditional modeling by means of logistic regression, as it allows for assessing homogeneity of preferences (choices) of the groups detected, instead of simply finding the key factors related to choice. We recommend combining regression analysis with decision trees modeling in all kinds of academic and applied research studying complicated choices.
Keywords: logistic regression, classification, choice modeling, conditional decision trees, school choice.

References

1. Breen R., Goldthorpe J.H. Explaining educational differentials: Towards a formal rational action theory, *Rationality and Society*, 1997, 9 (3), 275–305.
2. Ball S.J. Good school/bad school: paradox and fabrication, *British Journal of Sociology of Education*, 1997, 18 (3), 317–336.
3. Taylor C. Hierarchies and local markets: the geography of the lived market place in secondary education provision, *Journal of Education Policy*, 2001, 16 (3), 197–214.
4. Kristen C. *School Choice and Ethnic School Segregation: Primary School Selection in Germany*. Waxmann Verlag, 2003.

5. Shavit Y., Blossfeld H.P. *Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Social Inequality Series*. Boulder: Westview Press, 1993
6. Mare R.D. Change and stability in educational stratification, *American Sociological Review*, 1981, 1, 72–87.
7. Shavit Y., Blossfeld H.-P. *Persistent Inequality: Changing Educational Attainment in Thirteen Countries, Social Inequality Series*. ERIC, 1993.
8. Lucas S.R. Effectively Maintained Inequality: Education Transitions, Track Mobility, and Social Background Effects, *American Journal of Sociology*, 2001, 106 (6), 1642–1690.
9. Breen R., Jonsson J.O. Analyzing educational careers: A multinomial transition model, *American Sociological Review*, 2000, 65(5), 754–772.
10. Cullen J.B., Jacob B.A., Levitt S.D. The impact of school choice on student outcomes: an analysis of the Chicago Public Schools, *Journal of Public Economics*, 2005, 89 (5-6), 729–760.
11. Shaikhina T., Lowe D., Daga S., Briggs D., Higgins R., Khovanovaa N. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation, *Biomedical Signal Processing and Control*. 2017. https://doi.org/10.1016/j.bspc.2017.01.012.
12. Masias V.H., Valle M.A., Amar J.J., Cervantes M., Brunal G., Crespo F.A. Characterising the Personality of the Public Safety Offender and Non-offender using Decision Trees: The Case of Colombia, *Journal of Investigative Psychology and Offender Profiling*, 2016, 13 (3), 198–219.
13. Feldesman M.R. Classification trees as an alternative to linear discriminant analysis, *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 2002, 119, 257–275.
14. Karels T.J., Bryant A.A., Hik D.S. Comparison of discriminant function and classification tree analyses for age classification of marmots, *Oikos*, 2004, 105 (3), 575–587.
15. Guidotti R., Monreale A., Ruggieri S., Turini F. A Survey of Methods for Explaining Black Box Models, *ACM Computing Surveys (CSUR)*, 2018, 51 (5), 93.
16. Markou E. *3 Machine Learning Algorithms You Need to Know*. URL: https://dzone.com/articles/3-machine-learning-algorithms-you-need-to-know (date of access: 21.11.2018).

17. Morgan J.N., Sonquist J.A. Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association*, 1963, 58 (302), 415–434.
18. Breiman L., Friedman J.H., Olshen R.A., Stone C.J. *Classification and Regression Trees*. New York: Chapman and Hall. 1984.
19. Quinlan J.R. Induction of decision trees, *Machine Learning*, 1986, 1 (1), 81–106.
20. Quinlan J.R. *C4.5: Programms for Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc., 1993.
21. Strobl C., Malley J., Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests, *Psychological Methods*, 2009, 14 (4), 323–348.
22. Kingsford C., Salzberg S.L. What are decision trees? *Nature Biotechnology*, 2008, 26, 1011–1013.
23. Song Y.Y., Ying L.U. Decision tree methods: applications for classification and prediction, *Shanghai Archives of Psychiatry*, 2015, 27 (2), 130–135.
24. Hothorn T., Hornik K., Zeileis A. Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical statistics*, 2006, 15 (3), 651–674.
25. Zeileis A., Hothorn T. *partykit: A Toolkit for Recursive Partytioning*, 2012. URL: https://cran.r-project.org/web/packages/partykit/vignettes/partykit.pdf (date of access: 21.11.2018).
26. Hothorn T., Hornik K., Zeileis A. ctree: Conditional Inference Trees, *The Comprehensive R Archive Network*, 2015. URL: https://rdrr.io/rforge/partykit/f/inst/doc/ctree.pdf (date of access: 21.11.2018).
27. Hothorn T., Hornik K., Strobl C., Zeileis A. *Party: A Laboratory for Recursive Partitioning*, 2010. URL: https://cran.r-project.org/web/packages/party/vignettes/party.pdf (date of access: 21.11.2018).
28. Golland P., Liang F., Mukherjee S., Panchenko D. "Permutation tests for classification", in: *International Conference on Computational Learning Theory*. Springer, Berlin, Heidelberg, 2005, 501–515.
29. Therneau T.M., Atkinson E.J. *An Introduction to Recursive Partitioning Using the RPART Routines*, 2018. URL: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf (date of access: 21.11.2018).
30. Fabrikant M.S. Model-orientirovannyj podhod k otsutstvuyushchim znacheniyam: mnozhestvennaya imputaciya v mnogourovnevoj regressii

posredstvom R (na primere analiza oprosnyh dannyh) (in Russian), *Sotsiologiya 4M* (Sociology: Methodogy, Methods, Mathematical modeling), 2016, 41, 7–29.

31. Feelders A. "Handling missing data in trees: surrogate splits or statistical imputation?" in: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, Berlin, Heidelberg, 1999, 329–334.
32. Janssen K.J., Donders A.R. T., Harrell Jr. F.E., Vergouwe Y., Chen Q., Grobbee D.E., Moons K.G. Missing covariate data in medical research: to impute is better than to ignore, *Journal of Clinical Epidemiology*, 2010, 63 (7), 721–727.
33. Valdiviezo H.C., Van Aelst S. Tree-based prediction on incomplete data using imputation or surrogate decisions, *Information Sciences*, 2015, 311, 163–181.