УДК 378-057.87:316.344.34

Д.И. Сапонов

ОПЫТ КОНКУРЕНТНОЙ БОРЬБЫ КАК ФАКТОР АКАДЕМИЧЕСКОЙ УСПЕВАЕМОСТИ

ОПЫТ КОНКУРЕНТНОЙ БОРЬБЫ КАК ФАКТОР АКАДЕМИЧЕСКОЙ УСПЕВАЕМОСТИ

САПОНОВ Дмитрий Игоревич — преподаватель МВШСЭН. E-mail: Dsaponov1@yandex.ru.

Аннотация. В работе рассматриваются модели зависимостей текущей академической успеваемости студентов от опыта участия в различных видах социальной активности в недавнем прошлом (за последние три года).

Эмпирической базой исследования являются данные мониторинга образовательных и трудовых траекторий студентов и выпускников Школы педагогики Дальневосточного федерального университета.

Автор выделяет три вида социальной активности, значимые с точки зрения академической успеваемости (участие в олимпиадах, получение грамот за хорошую учебу и получение призовых мест на спортивных соревнованиях); показано, что недавний опыт социальной активности сильнее других факторов определяет текущую академическую успеваемость.

Предложены две классификационные модели для выявления взаимосвязей между академической успеваемостью и участием в конкурентных ситуациях: логистическая регрессия и дерево классификации. Для сравнения качества классификационных моделей был проведен ROC анализ. Для визуализации зависимости применяется анализ соответствий.

Ключевые слова: логистическая регрессия, дерево классификации, анализ соответствий, чувствительность, специфичность, ROC-анализ.

COMPETETIVE PRACTICES AS A FACTOR AFFECTING ACADEMIC PERFORMANCE SCORE

SAPONOV Dmitrii Igorevich – Lecturer, Moscow School of Social and Economic Sciences. E-mail: Dsaponov1@yandex.ru.

Abstract. This paper examines different models of dependencies of students` performance score on student involvement in different activities (for last three years).
The empirical basis of the study is the monitoring of education and career paths of students and graduates of the School of Pedagogy of the Far Eastern Federal University.

Three types of social activities are proposed by the author: participation in the Olympiads, getting diplomas for good studying and getting the winning places in sports competitions.
The recent social experiences dramatically affect current academic performance.

This paper suggests two classification models to reveal interrelationships between students` progress and their participation in competitive situations: logistic regression and classification tree. Classification models quality was evaluated using Receiver Operating Characteristic (ROC) analysis. Data were visualized using correspondence analysis.

Keywords: logistic regression, classification tree, correspondence analysis, sensibility, specifics, ROC analysis.

Данная работа выполнена по результатам инсталляции первой волны системы мониторинга образовательных и трудовых траекторий студентов и выпускников Школы педагогики Дальневосточного федерального университета (ДВФУ) в рамках проекта

«Консультационное и экспертное сопровождение деятельности руководства Школы педагогики ДВФУ в ходе разработки и запуска инновационной модели современного педагогического вуза (на базе Школы педагогики) как структурного подразделения классического (федерального) университета». Проект реализован в Институте образования НИУ ВШЭ. Исследователи выражают благодарность руководству ДВФУ и коллективу Школы педагогики ДВФУ за эффективное содействие в реализации проекта.

Основная задача работы — методологическая: поиск возможных способов описания и визуализации социологических данных. Сформулированные нами требования к способам представления данных продиктованы двумя соображениями. С одной стороны, способы представления данных должны быть достаточно просты и универсальны, для того чтобы применять их рутинным образом в любом социологическом исследовании. С другой стороны, необходимо продвинуться чуть дальше стандартных таблиц сопряженности, которые всегда были основным и максимально понятным интерфейсом, обеспечивающим пользователю доступ к информации, содержащейся в социологических данных.

База исследования

В исследовании принимали участие студенты разных ступеней образования и различных курсов (табл. 1). В соответствии с гипотезой о том, что студенты различных ступеней образования имеют существенно разную мотивационную структуру, далее мы будем исследовать влияние различных факторов на академическую успеваемость только для студентов 1-го курса бакалавриата. Это самая многочисленная группа (N=342).

Таблица 1 Структура данных по ступеням образования

	Частота	Процент
1-й курс бакалавриата	342	57,1
5-й курс специалитета	123	20,5
1-й курс магистратуры	104	17,4
2-й курс магистратуры	22	3,7
Итого	591	98,7
Системные пропущенные	8	1,3
Итого	599	100

Среди факторов, сильно связанных с академической успеваемостью, следует выделить следующие, существенно различные по характеру группы:

- фактический опыт; социальная активность за последние 3 года (дополнительное образование, спорт, олимпиады);
- субъективная самооценка (самоидентификация); субъективное позиционирование (согласие/несогласие) о специально подобранных высказываниях (например: «В работе я упорен (упорна)», «Я хорошо ужился (ужилась) с другими студентами»);
- ожидания; степень сформированности представлений о будущей работе, мотив выбора специальности («Обучение по этой специальности будет интересным»).

Остановимся на анализе влияния ранее приобретенного опыта на академическую успеваемость. На это есть две причины. Во-первых, показатели, связанные с недавним опытом социальной активности, сильнее других связаны с успеваемостью (табл. 2.) Как видно, первую и третью позиции занимают показатели, относящиеся к опыту социальной активности за последние 3 года.

Таблица 2 Показатели, сильнее остальных связанные с академической успеваемостью

Показатель	Статистическая значимость связи с академической успеваемостью
Участвовали в олимпиадах, соревнованиях или конференциях по математике, истории и другим предметам любого уровня	8,37338E-06
В работе я упорен (упорна)	5,28592E-05
Были награждены за успехи в учебе (грамотой, медалью и пр.)	1,12472E-04
Я хорошо ужился (ужилась) с другими студентами	1,31958E-04
Могли ли Вы к моменту поступления описать, что представляет работа по выбранной специальности?	3,15019E-04
Моя учеба носит бессистемный характер	4,73430E-04
Я чувствую поддержку со стороны других студентов	1,32120E-03
Я в хороших отношениях с преподавателями и сотрудниками вуза	1,68886E-03
Я часто чувствую, что был бы рад (была бы рада) покинуть вуз	2,79740E-03
Что Вы ожидаете от своей работы? Работа должна давать мне возможность распоряжаться рабочим временем	2,88911E-03

Во-вторых, эта группа показателей измеряется через регистрацию фактов: участие в мероприятиях, получение грамот, поощрений, призовых мест. Такая социологическая информация более достоверна по сравнению с самоидентификацией и ожиданиями [1].

Блок показателей, связанных с опытом социальной активности за последние 3 года, и распространенность каждого вида активности представлены на рисунке 1.

Рисунок 1 — Опыт социальной активности за последние три года

Самыми распространенными формами социальной активности являются получение грамот за успехи в учебе, участие в предметных олимпиадах, занятие творчеством (музыка, танцы, изобразительное искусство), спорт.

Итак, в качестве независимых переменных будем использовать виды социальной активности, в качестве зависимой (целевой) переменной будет выступать академическая успеваемость. Распределение переменной «Академическая успеваемость» представлено в таблице 3.

Таблица 3 Распределение переменной «Академическая успеваемость»

	Частота	Процент
Получаю высшие баллы по большинству предметов	45	13,2
Учусь хорошо, пересдач и хвостов практически не бывает	212	62,0
Учусь удовлетворительно, но бывают пересдачи и/или хвосты	76	22,2
Итого	333	97,4
Системные пропущенные	9	2,6
Итого	342	100,0

Логистические модели

Если использовать для моделирования зависимости логистическую регрессию, то зависимая переменная должна быть бинарной. В нашем случае из переменной «Академическая успеваемость» можно сконструировать 3 бинарные переменные, соответствующие трем категориям академической успеваемости. Каждая из этих бинарных переменных может быть использована в логистической регрессии в качестве зависимой переменной. В таблице 4 приведены результаты построения трех логистических моделей.

Таблица 4 Логистические регрессии, построенные для зависимых переменных, отражающих академическую успеваемость

	Получаю высшие баллы по большинству предметов			Учусь хорошо, пересдач и хвостов практически не бывает			Учусь удовлетворительно, но бывают пересдачи и/или хвосты		
	B	Знч.	Exp(B)	B	Знч.	Exp(B)	B	Знч.	Exp(B)
Участвовали в олимпиадах, соревнованиях или конференциях по математике, истории и другим предметам	1,05	0,02	2,85	0,35	0,18	1,42	–0,89	0,00	0,41
Были награждены за успехи в учебе (грамотой, медалью и пр.)	–0,14	0,71	0,87	0,60	0,02	1,83	–0,72	0,02	0,49

	Получаю высшие баллы по большинству предметов			Учусь хорошо, пересдач и хвостов практически не бывает			Учусь удовлетворительно, но бывают пересдачи и/или хвосты		
	B	Знч.	Exp(B)	B	Знч.	Exp(B)	B	Знч.	Exp(B)
Занимали призовые места на соревнованиях (лично Вы или Ваша спортивная команда)	–0,02	0,94	0,98	0,59	0,02	1,80	–0,73	0,01	0,48
Константа	–2,95	0,00	0,05	-0,07	0,79	0,93	-0,09	0,76	0,92

В результате построения логистических моделей из всего блока показателей опыта социальной активности были отобраны только статистически значимые: участие в предметных олимпиадах, получение грамот за успехи в учебе, завоевание призовых мест на спортивных соревнованиях. Так как база анализа небольшая (N=342), статистическая значимость коэффициента обязательно должна приниматься во внимание. В нашем случае небольшая база исследования позволила эффективно отобрать по критерию статистической значимости 3 важные независимые переменные. Построение логистических моделей позволяет сделать следующие выводы:

- для выделения группы «Троечники» значимы все 3 фактора: неучастие в предметных олимпиадах, неполучение грамот за хорошую учебу и неучастие в спортивных соревнованиях;
- для выделения группы «Хорошисты» значимы получение грамот за хорошую учебу и участие в спортивных соревнованиях;
- для выделения группы «Отличники» значимым фактором является только участие в предметных олимпиадах.

На первом шаге мы определили, какие показатели статистически значимо влияют на успеваемость в целом (олимпиады, грамоты за хорошую учебу, спортивные соревнования). На втором шаге мы хотели детализировать картину и понять, какие показатели позволяют выделить, например, «Троечников». Сопоставляя коэффициенты различных логистических моделей, мы установили, что «Троечников» определяют все три фактора, «Отличников» — только участие в олимпиадах, «Хорошистов» — получение грамот и завоевание призовых мест на спортивных соревнованиях.

Анализ соответствий как способ визуализации результатов

Анализ соответствий применяется для исследования таблиц сопряженности, он позволяет видеть столбцы и строки в виде точек, нанесенных на карту восприятия. В нашем случае входной таблицей для анализа соответствий будет таблица частот по академической успеваемости и трем независимым переменным (табл. 5). Точки строк и столбцов наносятся на одну и ту же карту и анализируются одновременно.

Таблица 5 Частотное распределение, подлежащее визуализации с помощью анализа соответствий

		Отличники	Хорошисты	Троечники
Участвовали в олимпиадах, соревнованиях или конференциях по математике, истории и другим предметам любого уровня	Да	38	146	33
	Нет	7	66	43
Были награждены за успехи в учебе (грамотой, медалью и пр.)	Да	33	156	38
	Нет	12	56	38
Занимали призовые места на соревнованиях (лично Вы или Ваша спортивная команда)	Да	21	105	23
	Нет	24	107	53

Метод имеет ряд особенностей, которые выгодно отличают его от других методов анализа таблиц сопряженности. Обычная процедура при анализе таблиц сопряженности — выявление меры общей связи (при помощи статистики χ^2), однако этот метод ничего не говорит о связях отдельных строк и столбцов. Анализ соответствий решает эту проблему.

Кратко анализ соответствий можно определить как особый случай метода анализа главных компонент строк и столбцов матрицы (таблицы сопряженности). Тем не менее анализ соответствий и анализ главных компонент используются при разных обстоятельствах. Метод главных компонент применяется для анализа непрерывных величин, в то время как анализ соответствий — в основном для анализа категориальных переменных. Основной целью анализа соответствий является графическое представление каждой строки и каждого столбца таблицы как точки на плоскости.

Для чтения карты восприятия удобно пользоваться следующими практическими правилами:

— Точки вблизи начала координат имеют малое влияние.
— Точки одного набора, расположенные далеко от центра, но близко друг от друга, имеют схожие профили.
— Геометрически отдельный профиль строки (высказывания) притягивается к такому положению на графике, где расположены столбцы, имеющие бросающиеся в глаза особенности в данной строке.

На рисунке 2 видно, например, что категория «Отличники» имеет один и тот же угол относительно начала координат, что и показатель «Участвовали в олимпиадах». Это говорит о том, что данный показатель сильнее остальных выделяет категорию «Отличники». Вместе с тем анализ соответствий, в отличие от логистической регрессии, не дает количественных оценок, он предназначен для визуальной оценки структуры взаимосвязей и выдвижения новых гипотез.

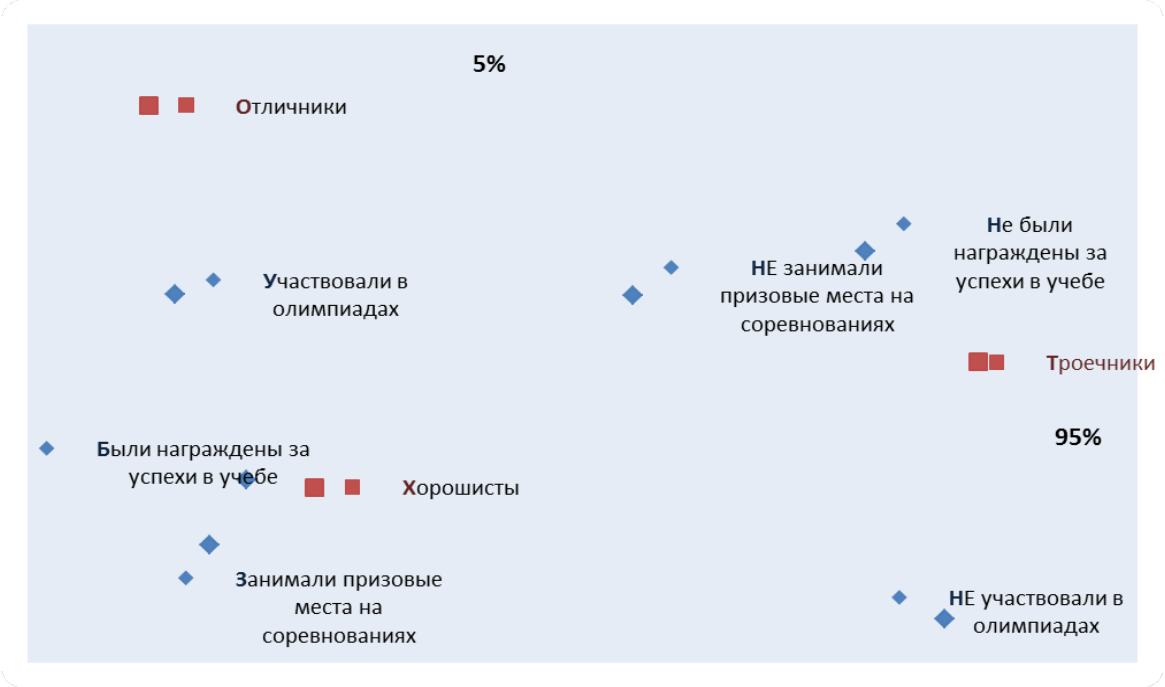

Рисунок 2 — Графическое представление результата анализа соответствий

Подписи на горизонтальной оси (95%) и на вертикальной оси (5%) указывают на вес осей в представлении информации. В нашем случае основная нагрузка ложится на горизонтальную ось, а дополнительное вертикальное измерение добавляет 5% информации.

Классификационная модель

Главная особенность классификационных алгоритмов состоит в том, что они формируют модель не в аналитическом виде, а в виде набора правил, по которым определяется вероятность принадлежности к той или иной группе. Например, в отсутствии модели вероятность попадания респондента в категорию «Троечники» составляет 22,8%. В соответствии с классификационной моделью для тех, кто не участвовал в олимпиадах и не был награжден грамотой за успехи в учебе, вероятность попадания в категорию «Троечники» возрастает с 22,8 до 48,3%.

В моделях дерева решений выполняется последовательное разбиение набора данных на основе взаимосвязи между предикторными переменными и целевой (результирующей) переменной. Полученное в результате дерево показывает, какие независимые переменные наиболее тесно связаны с целевой переменной. Также выводятся подгруппы (конечные узлы), в которых могут концентрироваться наблюдения, имеющие желаемые характеристики (рис. 3).

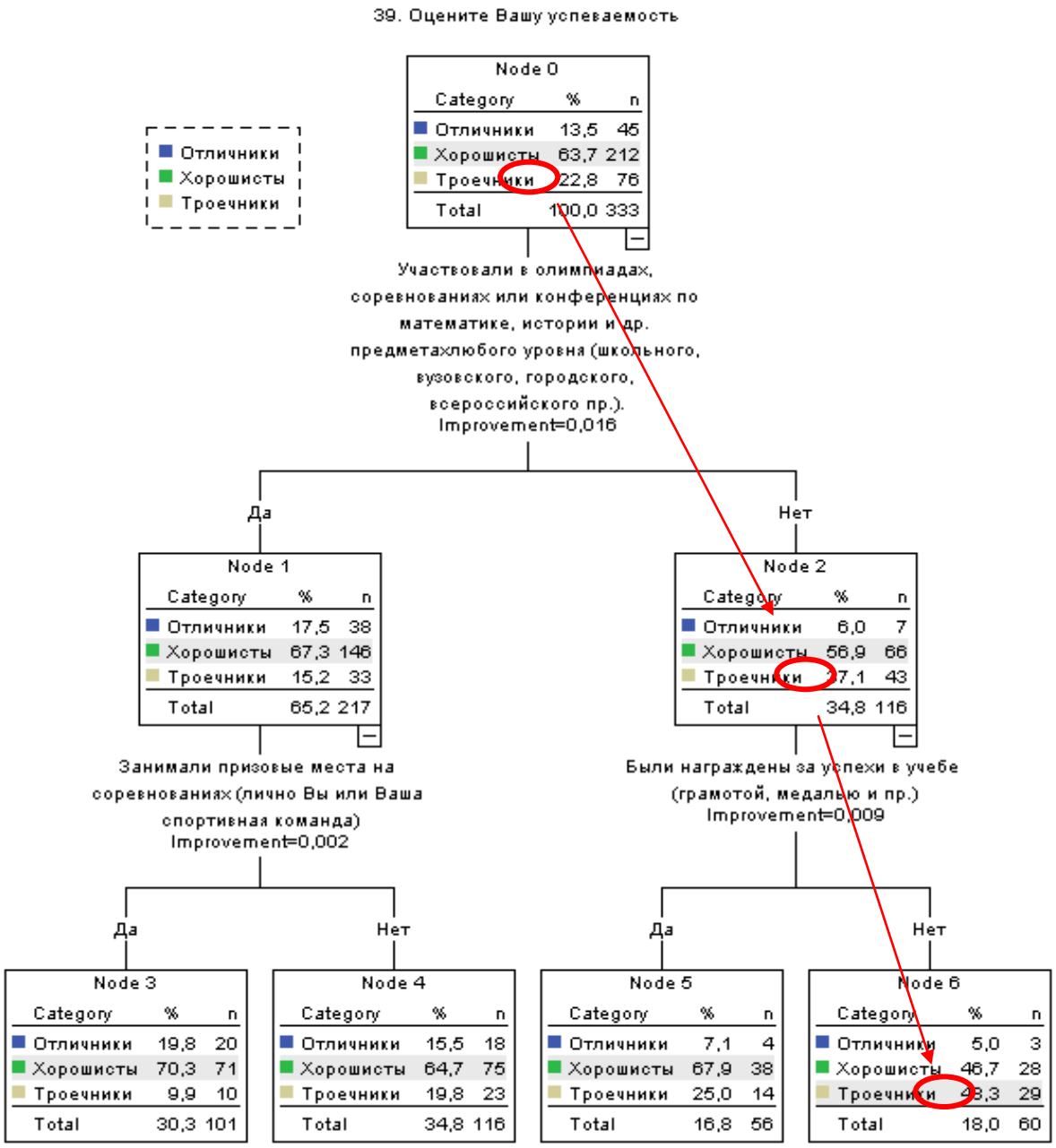

Рисунок 3 — Классификационная модель академической успеваемости

Сравнение качества моделей. ROC-анализ

В заключение проведем сравнение качества моделей логистической регрессии и дерева классификации. Сравним модели, построенные для зависимой переменной «Учусь удовлетворительно, но бывают пересдачи и/или хвосты», значения 0 и 1.

Для оценки качества моделей с бинарной зависимой переменной применяется таблица классификации. Фактически, это таблица сопряженности по исходной целевой переменной (в нашем случае это «Учусь удовлетворительно, но бывают пересдачи и/или хвосты», значения 0 и 1) и по предсказанному моделью значению этой переменной. Чем

выше совпадение переменных, тем выше качество модели. Обычно подсчитывается процент правильно предсказанных значений в каждой группе.

Как правило, для сравнения качества классификационных моделей достаточно сравнить две таблицы классификации — для логистической регрессии и дерева классификации представлены (см. табл. 6 и 7).

Таблица 6 Таблица классификации. Логистическая регрессия

		Предсказанные		
		учусь удовлетворительно, но бывают пересдачи и/или хвосты		процент корректных
		0	1	
Учусь удовлетворительно, но бывают пересдачи и/или хвосты	0	260	0	100,0
	1	76	0	0,0
Общий процент				77,4
Разделяющее значение = 0,5				

Таблица 7 Таблица классификации. Дерево классификации

Наблюденное	Предсказанное		
	,00	1,00	Процент корректных
,00	260	0	100,0%
1,00	76	0	0,0%
Общий процент	100,0%	0,0%	77,4%

Так как целевая группа (те, кто учится удовлетворительно, но бывают пересдачи и/или хвосты) составляет порядка 22% от всего массива данных, предсказанное значение смещено в сторону более наполненной остаточной группы. В результате обе модели не дают ни одного предсказанного значения для целевой группы, что видно из таблиц классификации.

Для более детального анализа нужно рассмотреть все множество таблиц классификации, соответствующих разным порогам классификации. Порог классификации — это значение вероятности, которое является разделяющим при решении о вычислении принадлежности к целевой группе. В моделях, которые были построены, порог классификации (разделяющее значение) равен 0,5, т.е. если предсказанная вероятность более 0,5, модель относит наблюдение к группе 1. В наших моделях ни одно предсказанное значение вероятности не перешло за порог 0,5, что и привело к одинаковым таблицам классификации для разных моделей. Но, из того что при пороге классификации 0,5 модели ведут себя одинаково, не вытекает, что модели идентичны. Выходом из этой ситуации, который позволит сравнить две модели и увидеть различия в их характеристиках, будет сравнение множества таблиц классификации при варьировании порога классификации. Для наглядного представления результатов этой процедуры используются так называемые ROC-кривые или кривые ошибок, а оценка моделей с помощью ROC-кривых называется ROC-анализом [3, p. 232].

Термин «операционная характеристика приемника» (Receiver Operating Characteristic, ROC) восходит к теории обработки сигналов. Эта характеристика впервые была предложена во время Второй мировой войны, после поражения американского военного флота в Пёрл-Харборе в 1941 г., когда возникла проблема повышения точности распознавания самолетов противника по радиолокационному сигналу. Сейчас ROC-анализ широко применяется и в других сферах: при медицинской диагностике, контроле качества, кредитном скоринге, предсказании лояльности клиентов и т.д.

Одна точка на кривой ошибок соответствует одной таблице классификации. ROC-кривая соответствует разным таблицам классификации при различных порогах классификации. Фактически на кривой по горизонтальной оси откладывается значение B, выраженное в процентах по строке (1 — специфичность), а по вертикальной оси — значение A (чувствительность), тоже выраженное в процентах по строке (табл. 8).

Таблица 8 Выражение координат ROC-кривой через значения в ячейках таблицы классификации

		Предсказанные		
		0	**1**	**ВСЕГО**
Наблюдаемые	0	A (Чувствительность)	C	A+C
	1	B (1 — Специфичность)	D (Специфичность)	B+D
		A+B	C+D	A+B+C+D
Чувствительность = A/(A+C) Специфичность = D/(B+D)				

На плоскость можно нанести несколько кривых, соответствующих разным моделям. На рисунке 4 изображены две ROC-кривые, соответствующие двум моделям, описанным выше (логистическая регрессия и дерево классификации).

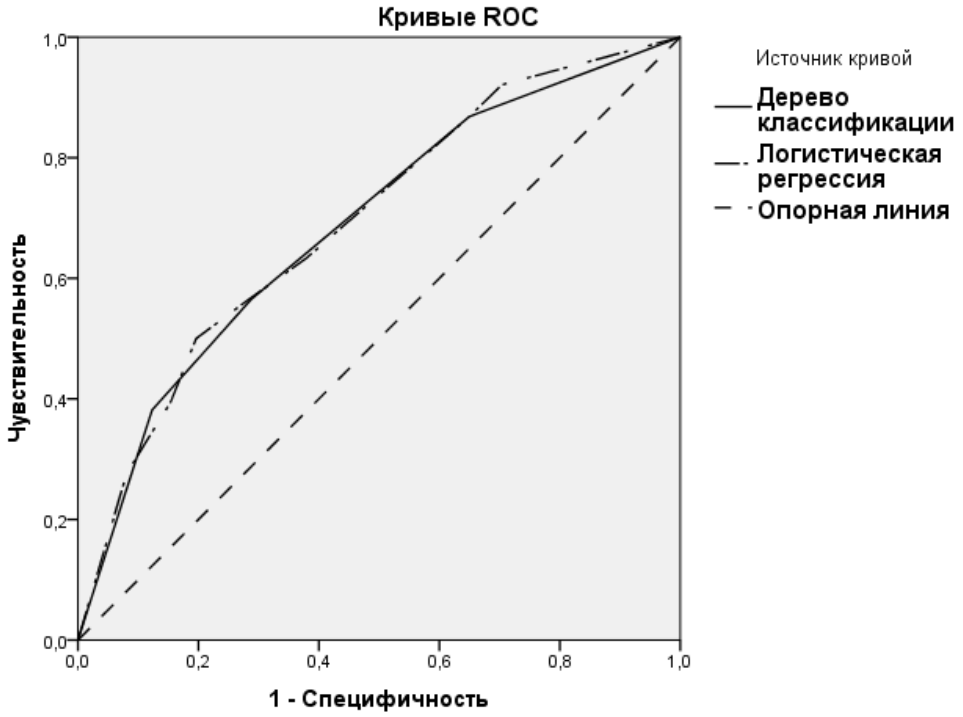

Рисунок 4 — ROC-кривые для двух моделей классификации

Анализ ROC-кривых позволяет решить две задачи:

1. Выбрать порог классификации таким образом, чтобы сумма правильно предсказанных значений была максимальна [2, p. 1701]. Для этого можно воспользоваться координатами ROC-кривой (табл. 9). Сумма правильно предсказанных соответствует сумме (чувствительность + специфичность). Для дерева классификации оптимальный порог классификации равен 0,22, для логистической регрессии — 0,30.

Таблица 9 Координаты ROC-кривой

Тестовая переменная(ые)	Положительное, если больше или равно	Чувствительность	1 — Специфичность
Дерево классификации	,0000	1,000	1,000
	,1478	,868	,650
	,2211	,566	,288
	,3605	,382	,123
	1,0000	,000	,000
Логистическая регрессия	,0000000	1,000	1,000
	,1270457	,921	,704
	,1639743	,868	,650
	,1874120	,632	,377
	,2424535	,566	,285
	,3061324	,500	,196
	,3392595	,395	,154
	,4205531	,276	,081
	1,0000000	,000	,000

2. Сравнить две модели между собой и выбрать лучшую. Лучшей считается модель, у которой больше площадь под кривой (табл. 10). В нашем случае площадь несколько больше под кривой, соответствующей логистической регрессии, однако анализ доверительных интервалов для площадей под кривыми показывает отсутствие значимых различий, т.е. в нашем примере прогноза количества «Троечников» обе модели работают одинаково хорошо.

Таблица 10 Площадь под ROC-кривой

Тестовая переменная(ые)	Площадь	Стд. ошибка a	Асимптотическая знч. b	Асимптотический 95%-ный доверительный интервал	
				нижняя граница	верхняя граница
Дерево классификации	,688	,035	,000	,620	,757
Логистическая регрессия	,694	,034	,000	,627	,761
b. Нулевая гипотеза: истинная площадь = 0.5					

Обсуждение результатов

Выбор академической успеваемости в качестве целевой переменной для построения моделей обусловлен широким контекстом исследования жизненных траекторий учащихся. Академическая успеваемость, с одной стороны, является легко измеряемым устойчивым объективным показателем, с другой стороны — представляет сложный комплекс, аккумулирующий множество мотивационных, стилевых, биографических, стратификационных параметров. Из прошлого успеваемость определяется социальным статусом семьи и личным опытом социальной активности, будущее определяет академическую успеваемость через мотивацию, ожидания и амбиции.

Можно предположить, что во время обучения в университете по мере врастания в профессиональную жизнь академическая успеваемость как показатель усложняется, становится более дифференцированной, постепенно замещаясь общей профессиональной состоятельностью и переставая быть столь доступной для измерения. Таким образом, согласно нашему предположению, академическая успеваемость является сравнительно легко измеримым на младших курсах прототипом профессиональной состоятельности, поэтому в части эмпирических выкладок мы остановились только на группе студентов 1-го курса бакалавриата.

При анализе массива опросных данных были выделены 3 группы показателей, сильно связанных с академической успеваемостью: опыт социальной активности за последние 3 года, субъективная идентификация, ожидания. Для анализа были отобраны показатели, связанные с опытом социальной активности по двум причинам. Во-первых, эта группа показателей сильнее остальных определяет академическую успеваемость. Во-вторых, эти показатели являются данными о фактическом поведении, что делает их более достоверными.

Из всего блока показателей, связанных с социальной активностью, значимыми оказались 3 показателя: участие в олимпиадах, получение грамот за хорошую учебу и получение призовых мест на спортивных соревнованиях. Эти показатели так или иначе связаны с конкурентной ситуацией. Поэтому важным результатом эмпирического анализа

является вывод, что наличие опыта конкурентной борьбы увеличивает шансы на хорошую академическую успеваемость.

Далее мы наметили пути дальнейшей детализации модели, которая, на наш взгляд, должна развиваться по пути выделения двух относительно небольших групп: «Отличников», которые составляют 13%, и «Троечников», которые составляют 22% опрошенных. Мы наметили некоторую дифференциацию показателей наличия опыта конкурентной борьбы. Так, попадание в группу «Отличники» определяется только показателем «Участие в олимпиадах», а попадание в группу «Троечники» в равной степени определяется всеми тремя показателями.

Для наглядного представления результатов и дополнения логистической модели был проведен анализ соответствий, результаты которого представлены в виде карты восприятия, позволяющей визуально оценить характер взаимосвязей. Анализ соответствий в целом согласуется с дифференциацией входных показателей, которую мы наметили по результатам логистической модели. Отметим, что сильная сторона анализа соответствий связана с возможностью визуализации больших таблиц. В нашем случае мы имели дело с небольшой таблицей 3×6 (3 категории успеваемости и 3 показателя по 2 категории в каждом). Если значимых показателей, определяющих успеваемость, будет больше, анализ соответствий выступит как более мощный инструмент, позволяющий одновременно оценивать совокупность взаимосвязей.

Наконец классификационная модель одновременно представляет результат в виде наглядного и интуитивно понятного дерева решений и дает возможность прогнозировать целевую переменную.

Особенности различных методов анализа, которые мы использовали, приведены в таблице 11.

Таблица 11 Специфика использованных методов анализа

	Возможность прогноза целевой переменной	**Прогноз в аналитическом виде (формула)**	**Наглядность результата**	**Возможность одновременного визуального оценивания большого числа показателей**
Логистическая регрессия	Да	Да	Нет	Нет
Анализ соответствий	Нет	Нет	Да	Да
Деревья классификации	Да	Нет	Да	Нет

Литература

1 Шляпентох В. Э. Проблемы качества социологической информации: достоверность, репрезентативность, прогностический потенциал. — М. : ЦСП, 2006.

2 Elswick R. K., Schwartz P. S., Welsh J. A. Interpretation of the odds ratio from logistic regression after a transformation of the covariate vector // Statistics. 1995. Vol. 16. P. 1695–1703.

3 Pearce J., Ferrier S. Evaluating the predictive performance of habitat models developed using logistic regression // Ecological Modelling. 2000. Vol.133. P. 225–245.