ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет бизнеса и менеджмента

АНАЛИЗ И ПРОГНОЗИРОВАНИЕ АКАДЕМИЧЕСКОЙ УСПЕВАЕМОСТИ СТУДЕНТОВ НА ПРИМЕРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ «БИЗНЕС-ИНФОРМАТИКА» НИУ ВШЭ

Васюкова Виктория Андреевна

3 курс, направление подготовки: 38.03.05 «Бизнес-информатика»
образовательная программа «Бизнес-информатика»

Научный руководитель
Е. С. Прокофьева

Москва, 2020

СОДЕРЖАНИЕ

1. ВВЕДЕНИЕ

Академическая успеваемость студентов является важным показателем работы образовательного учреждения, отражающим отчасти вовлеченность студентов в образовательный процесс. Данная работа посвящена одной из ее составляющих — отчислениям — и фокусируется на успеваемости студентов Национального исследовательского университета «Высшая школа экономики» (далее НИУ ВШЭ). Ограничения, накладываемые форматом доступных данных, не позволяют охватить сразу все образовательные программы, поэтому в анализе задействована только одна: рассматриваются академические исходы студентов программы бакалавриата «Бизнес-информатика» НИУ ВШЭ с использованием рейтинговых таблиц за 2014–2019 учебные годы, размещаемых на официальной странице [1] образовательной программы (далее ОП). Основная **цель** работы — построить модель отчислений студентов. Реализация подобной модели также позволит выявить причины отчислений и сформировать рекомендации для восстановления или улучшения успеваемости студентов на грани выбытия. **Актуальность** данной темы определяется тем, что от большого числа прерванных образовательных траекторий страдают как непосредственно студенты, так и, неся определённые экономические потери, сами образовательные учреждения. В соответствии с поставленной целью исследования были определены следующие **задачи**:

- обзор и глубокий анализ литературы
- сбор необходимых для исследования данных и приведение их к нужной структуре
- разведочный анализ полученных данных: оценка масштабов отсева, число пересдач, средняя успеваемость, гендерный состав учащихся
- анализ академической успеваемости студентов, а именно — рейтингов (подробнее о них будет сказано в разделе Данные)
- построение логистических моделей, деревьев решений и использование метода опорных векторов для прогнозирования отчислений и определение факторов, влияющих на отсев студентов

Работа имеет следующую структуру: в главе 2 приведён обзор имеющихся работ по данной тематике, глава 3 посвящена описанию сбора и обработки данных, а главы 4 и 5 — моделям прогноза.

2. ОБЗОР ЛИТЕРАТУРЫ

2.1. ОПРЕДЕЛЕНИЕ ОТСЕВА

Определение студенческого отсева в российских учебных заведениях предлагается в статье [2] как альтернатива зарубежному *dropout*, которым принято обозначать добровольное выбытие студента. Отсев в контексте российского образовательного процесса означает «принудительное по отношению к студенту действие, субъектом которого является университет». Отсутствие точных данных о масштабах отчислений из российских вузов делает проблематичным исследование такого важного феномена. В то же время Организация экономического сотрудничества и развития оценивает отчисляемость в 21% от численности обучающихся для России [3]. Существует мнение, что в России большой отсев студентов связывается с престижностью вуза, в то время как в западных странах большой процент отчислений говорит о неэффективности работы института. Это подтверждается статистикой [4] таких элитарных университетов как The University of Oxford и The University of Cambridge, где процент отчислений не превышает 1,5%.

2.2. ФАКТОРЫ, ВЛИЯЮЩИЕ НА АКАДЕМИЧЕСКИЙ ИСХОД

Так как академическому исходу так или иначе предшествует академический процесс и некая успеваемость, справедливо предполагать, что одной из первоочередных характеристик студентов, интересующих исследователей, является оценка навыков в разные промежутки времени, особенно на входе. Уже многие годы главной оценкой школьных знаний предстают результаты Единого государственного экзамена (ЕГЭ). Несмотря на споры о состоятельности ЕГЭ как инструмента такой оценки, ряд работ говорит о значимом влиянии баллов ЕГЭ на отметки, полученные в вузе. В среднем, результаты ЕГЭ объясняют 25–30% вариации успеваемости студентов при получении уже высшего образования [5].

В статье [6] исследуется взаимосвязь между различными показателями (в первую очередь отношением к риску) и вероятностью отчисления из вуза. Строятся несколько регрессионных моделей на выборке из студентов, зачисленных в высокоселективный вуз (вуз с жестким отбором на входе и высоким процентом отчислений в случае возникновения академической задолженности в процессе обучения) в 2010 г. на программы четырёх факультетов. Незначимыми оказываются такие факторы, как пол студента и материальное положение семьи. Влияние на текущую академическую успеваемость оказывают общая склонность к риску (негативное), результаты

ЕГЭ по русскому языку (положительное). Вероятность быть отчисленным связана с теми же факторами противоположным образом.

В работе, посвященной анализу гендерных различий в выбытии из вуза [7], пол студента выделяют как значимую в определении вероятности быть отчисленным характеристику лишь до включения в регрессионную модель достаточного количества других факторов. Несмотря на то, что статистика свидетельствует о преимуществе у девушек (95% девушек остаются в российских вузах к концу весеннего триместра первого курса в противовес 89% юношей; для американских студентов вузов Огайо цифры отличаются не так разительно — 82 и 81% для юношей и девушек соответственно), при включении в модель переменных, отражающих успеваемость в школе, группу специальностей обучения и среднюю оценку за первый семестр, а также факта проживания в общежитии, пол перестаёт быть значимым показателем на уровне значимости меньше или равном 0,05. Основными же факторами в определении уровня отчисления становятся специальность и академические успехи за первый семестр, а наибольшее выбытие из вуза приходится на первые два года обучения.

2.3. ИССЛЕДОВАНИЯ АКАДЕМИЧЕСКОЙ УСПЕВАЕМОСТИ РОССИЙСКИХ ВУЗОВ

В исследовании отчислений на примере Московского государственного университета [8], помимо уже рассмотренной нами проблемы определения факторов, влияющих на результат обучения, поднимается вопрос о временных и экономических потерях, связанных с преждевременным прерыванием обучения. При проведении исследования были использованы административные данные шести факультетов МГУ, а именно информация о студентах, зачисленных в 1998–2002 гг., также были привлечены результаты анализа когорт 1993–1997 гг. поступления. Работа вновь подтверждает тезис о том, что вероятность отчисления на первом курсе максимальна и достигает минимума к концу обучения. Здесь же делается замечание о том, что иная форма прерывания обучения — академический отпуск — используется студентами с одинаковой частотой на протяжении первых трех курсов. Отчисленных по своему желанию студентов 18–34% от общего числа. Из-за академической «неуспеваемости» уходят 45–59%, что значительно больше. Успешность обучения дифференцируется по полу — к группе высокого риска относятся юноши-студенты 1–2 курсов, это ставит под сомнение описанный ранее вывод о том, что в конечном итоге пол не играет ключевую роль в определении академического успеха. Ещё одной значимой характеристикой в исследовании видится возраст поступившего лица — высок риск отчисления

у студентов, поступивших в 21 год и старше; поступившие сразу после окончания школы достигают больших успехов. Последним важным фактором выделяется место жительства до начала обучения. Удельный вес выбывших более высок среди иногородних, что объясняется «двойной адаптацией» на первом курсе — одновременно к новому месту обучения и самостоятельной жизни. И всё-таки, можно ли ориентироваться на результаты данного исследования, если оно проводилось до введения ЕГЭ в качестве основного инструмента отбора при поступлении в вуз? В любом случае заключительная часть статьи может помочь яснее аргументировать актуальность исследований студенческих отчислений. Из раздела о возможных экономических потерях, к коим причисляются нерационально вложенные инвестиции в обучение в вузе, упущенные доходы за период обучения и доходы последующих лет, очевиден вывод о том, что потери от прерывания обучения существенны. Средние потери инвестиций, вложенных в обучение студентов, не закончивших, например, экономический факультет МГУ оценены в 2001 году в 4 300 тыс.руб. при обучении на бюджете и 15 360 тыс.руб. при обучении по контракту.

Тем не менее, наиболее интересными являются статьи, детально рассматривающие ситуацию в изучаемом вузе – НИУ ВШЭ. Отдельно стоит отметить ряд работ, посвященных анализу успеваемости студентов Международного института экономики и финансов (МИЭФ). Несмотря на то, что бакалаврская программа института специфична, политика приема абитуриентов и структура образовательного процесса те же, что и у остальных подразделений вуза.

В одной из таких работ [9] рассматривается зависимость академических успехов студентов, поступивших в МИЭФ в 2009– 2011 гг., после 1–3-го года обучения от достижений по ЕГЭ и олимпиадам. Строятся регрессионные модели с рейтингами студентов в качестве зависимой переменной, а также *logit*-модели вероятности выбывания студента. Результаты ЕГЭ по всем предметам (русский язык, математика, английский язык) оказываются значимы на 1%-ном уровне для прогноза рейтинга 1-го года, причем ЕГЭ по математике важен больше остальных. Поступившие по результатам Всероссийской олимпиады школьников, при прочих равных, получают рейтинги на 16–18 баллов выше. Что касается *logit*-модели, её результаты аналогичны результатам модели с рейтингом: высокие баллы ЕГЭ снижают вероятность быть отчисленным. Очевидно, что начальные данные о способностях студентов аккумулируются в их текущих достижениях — рейтинге и баллах по отдельным предметам. После включения в регрессоры рейтинга 1-го курса начальные данные — результаты ЕГЭ и олимпиад —

становятся незначимыми как в моделях рейтингов, так и в модели вероятности выбытия на 2–3 курсах. Интересно, что незначимым во всех моделях остается регион окончания школы. Однако оказывается значимой роль пола студента — при прочих равных студентки учатся лучше.

Следующая рассмотренная статья [10] исследует факторы достижений студентов МИЭФ на внешних экзаменах (международные экзамены проводятся Лондонским университетов в конце года). Упор делается на академические результаты по курсу эконометрики. В работе используются модели множественной регрессии и модели двоичного выбора. В качестве возможного показателя, интегрирующего способности, важные для прохождения экзаменов, выделяется успешность на предшествующих экзаменах (*proxy variable*). То есть промежуточные итоги в течение года включают в себя и уровень начальной подготовки, и уровень усвоения нового материала, и отношение студента к оценкам. Оценивание ряда моделей показывает, что различные показатели, отражающие выполнение домашних заданий, незначимы — это средняя оценка за домашние задания в течение года и отдельно по семестрам, число выполненных заданий. Это может быть объяснено следующими факторами: навыки, проверяемые в домашних работах, также проверяются на экзаменах; студенты склонны компенсировать недостаточную активность при выполнении домашних работ подготовкой к другим формам контроля, то есть к экзаменам; домашние задания могут включать в себя задачи, для выполнения которых требуется использование программного обеспечения, владение которым не проверяется в задачах экзамена. При этом в качестве одного из регрессоров может выступать посещаемость лекций — регулярно присутствовавшие на лекциях лица получили на апрельском экзамене МИЭФ примерно на 10 пунктов больше, чем регулярно не присутствовавшие. Полученные в итоге модели характеризуются высокой объясняющей способностью, изученные зависимости могут быть использованы для выявления расхождений между результатами обучения в НИУ ВШЭ и баллами за независимые экзамены, нахождения так называемых «проблемных» курсов.

Целью ещё одной работы по данной тематике [11] была попытка установить связь между успехами студентов МИЭФ с 2009 по 2013 год и политикой приема абитуриентов, а именно назначением скидок, распределением по группам, а также *peer effects*, то есть влиянием окружающих людей. В ходе исследования были использованы *regression discontinuity design* для определения роли размера скидки на коммерческой форме обучения и эффекта распределения по сильным и слабым академическим группам; стандартный МНК и бинарные модели для изучения

разброса результатов внутри групп. В результате автор приходит к следующим выводам: наличие или отсутствие скидки, определение студента в группы высокого уровня значимо не влияет на успеваемость; чем более однороден уровень группы, тем лучше результаты студентов; основной эффект от окружения испытывают на себе студенты в середине распределения, для них важен средний уровень группы и чем он выше, тем тяжелее им успевать за остальными, однако здесь делается поправка на специфику обучения в МИЭФ — студенты большую часть времени занимаются индивидуальной работой, семинарские занятия редки, из-за чего редки и взаимодействия между одногруппниками.

2.4. МЕТОДЫ ПРЕДСКАЗАНИЯ АКАДЕМИЧЕСКОЙ УСПЕВАЕМОСТИ

Систематическое исследование работ, фокусирующихся на прогнозе отчислений студентов [12], даёт следующую картину:

- примерно 79% работ используют легко интерпретируемые деревья решений, так как приходится иметь дело с данными разной природы – как численными, так и категориальными
- популярность деревьев решений оправдывается их точностью: классификаторы C4.5, ID3 и CART дают 98, 97,5 и 97% соответственно
- точность линейной регрессии – 87,8%
- возраст, пол, национальность и результаты вступительных испытаний являются самыми часто используемыми факторами для построения предсказательной модели

Конкретный пример использования описанных методов можно найти в работе, посвященной анализу немецких студентов [13]. С помощью методов машинного обучения строится предиктивная модель выбытия, задача которой состоит в раннем обнаружении потенциально отчисляемых студентов. Административные данные двух университетов включают в себя как демографические характеристики учащихся (пол, возраст, адрес, место рождения, сведения о миграции, предыдущие места и результаты учебы, курс и форма обучения), так и их академические достижения в каждом семестре (средняя оценка, число полученных кредитов, число пропущенных и не сданных экзаменов). Вводится переменная, обозначающая количество «самых важных» экзаменов, которые сдал студент. Они определяются по корреляции между сдачей и окончанием обучения у предыдущих когорт.

В работе используется алгоритм AdaBoost, позволяющий объединить предсказательные силы нейронной сети, регрессионной модели и бэггинга в один сильный классификатор, увеличив тем самым точность предсказаний. Важным условием для использования такого алгоритма является включение в

него действительно работающих методов, которые изначально имеют одинаковый вес, модифицирующийся на каждой итерации работы алгоритма. Касательно методологии делается замечание о том, что успешность исследования в большей мере зависит от предсказательной силы данных, нежели от используемого метода. Точность предсказания используемых моделей растёт со временем: например, для *probit*-модели доля идентифицированных находящихся под угрозой отчисления студентов в первом семестре равна 71%, в четвертом — уже 79%; для AdaBoost этот показатель достигает примерно тех же значений. Судя по приведённым в статье ROC curve для всех применяемых методов, нейросети и *probit*-модели уступают в точности BRF *(bagging random forest*), причём *probit* отстают значительно меньше. Также для обоих университетов верно, что демографические данные вносят меньший вклад в предсказательную способность модели, чем академическая успеваемость. Более того, они становятся всё менее значимыми при добавлении новых данных об успеваемости.

3. ДАННЫЕ

3.1. ОПИСАНИЕ И ОБРАБОТКА

Итак, в литературе на данную тему достаточно полно описаны механизмы и причины отсева студентов, связанные с их включенностью в социальную среду университета, социально-демографическими характеристиками и успеваемостью. В данной работе были рассмотрены лишь некоторые из них. Это объясняется отсутствием доступа к полным административным данным о студентах.

Рейтинговые таблицы, о которых было сказано ранее, регулярно публиковались в одном и том же виде с 2015 по 2018 годы. Рейтинг в контексте образовательного процесса в НИУ ВШЭ — это упорядоченный по нормированной сумме произведений оценок на кредитные веса дисциплин список студентов. Рейтинговая система позволяет ранжировать студентов по их успехам в учёбе и используется для определения скидки, которую может получить студент, обучающийся на коммерческом месте; участия студента в конкурсе на программу академической мобильности или повышенную стипендию и т.д. Помимо места студента среди однокурсников из рейтинговых таблиц известны средний и минимальный баллы, наличие академических задолженностей и оценки по всем пройденным предметам. Рейтинги публикуются до и после пересдач каждый семестр и бывают текущими и кумулятивными (такие рейтинги включают в себя оценки за проектную работу и факультативы).

Однако, уже после первого семестра 2018-2019 учебного года рейтинги были переведены из формата xlsx в формат небольшой таблицы на сайте, лишенной множества нужных для данного исследования характеристик учащихся. Так, с 2018 года в открытом доступе имеются лишь имена, позиции в рейтинге, средний балл, минимальный балл и GPA. Это всё еще позволяет определить студентов, идущих на пересдачу (минимальная оценка ниже 4), но проблемные дисциплины и количество пересдач остаются скрытыми — их можно определить, лишь обрабатывая разрозненные формы для записей на пересдачи. В свою очередь из собранных за 2015–2018 годы данных не удаётся узнать вид места, на котором обучаются студенты, и упускаются такие факторы, как баллы ЕГЭ и наличие скидки на обучение (подробная информация о зачисленных ежегодно обновляется). Несмотря на то, что результаты ЕГЭ остаются без внимания, следующий ниже анализ основывается на предположении о том, что школьная успеваемость отражается в рейтинге за первый семестр первого курса.

Результаты сбора и предобработки данных о первокурсниках, зачисленных в 2015, 2016 и 2017 году, частично представлены в следующей таблице. Помимо указанных в ней переменных, в анализе были задействованы оценки по 11 предметам, включенным в учебный план всех рассматриваемых первокурсников. Для визуализации динамики повторных поступлений (Рис. 4) были собраны данные о зачисленных в 2018–2019 годы.

Таблица 1. Соответствие колонок в исходных таблицах и названий переменных для анализа

id	уникальный идентификатор студента
gender	пол (1 - муж, 0 - жен)
min_score	минимальная оценка за семестр
avg_score	средняя оценка за семестр
rank	кредитно-рейтинговая сумма за семестр
fails	наличие неудовлетворительных оценок и неявок (1 - да, 0 - нет)
dropout	отчислен (1 - да, 0 - нет)
enrol_year	год зачисления

В общем случае названия переменных построены по одному шаблону: через нижнее подчеркивание («_») добавляются номер курса, семестра и «*before*» или «*after*» для маркирования периода до или после пересдач соответственно. Например, *avg_score_1_1_before* содержит данные средних оценках за первый семестр первого курса до пересдач, а *phil_after* — оценки по философии после пересдач. Переменная *gender* была извлечена из ФИО студентов. Отчества (при отсутствии использовалось имя) студентов были обработаны с помощью морфологического анализатора pymorphy2, который с высокой точностью способен определить род.

Таким образом, итоговый датасет, агрегирующий информацию о первых курсах 2015–2017 годов, состоит из 687 наблюдений и 49 переменных, включая фиктивные для определения года зачисления и периода выбытия.

3.2. РАЗВЕДОЧНЫЙ АНАЛИЗ

Гендерный состав когорт представлен в таблицах 2 и 3. Юношей больше как среди поступивших, так и среди отчисленных студентов.

Таблица 2. Пол студентов

Год поступления	**Пол**	
	муж	**жен**
2015	57,1%	42,9%
2016	57,4%	42,6%
2017	65%	35%

Таблица 3. Пол отчисленных

Год поступления	**Пол**	**Число отчисленных, чел.**
2015	жен	22
	муж	29
2016	жен	21
	муж	30
2017	жен	12
	муж	31

В таблице 4 отображены масштабы пересдач. Число пересдач равно количеству человек, получивших неудовлетворительную оценку по предмету и вынужденных сдать экзамен повторно для получения 4 баллов в итоге. Максимальные в процентном выражении цифры из года в год принадлежат математическим дисциплинам: Линейной алгебре, Математическому анализу и Дискретной математике. На протяжении всех включенных в анализ лет минимальное число студентов получает неудовлетворительные оценки по дисциплине Безопасность жизнедеятельности — менее 1% от числа поступивших.

Что касается масштабов отчислений, в 2015 и 2016 году было отчислено по 51 человеку, в 2017 — 43. Учитывая, что число первокурсников каждый год растёт, в процентном отношении отчисляемых становится меньше (Рис. 1).

Таблица 4. Пересдачи по предметам

Дисциплина	2015		2016		2017	
	Число пересдач	% от зачисленных	Число пересдач	% от зачисленных	Число пересдач	% от зачисленных
Безопасность жизнедеятельности	0	0,00	2	0,87	0	0,00
История	5	2,30	46	20,00	4	1,67
Линейная алгебра и геометрия	57	**26,27**	64	**27,83**	73	**30,42**
Математический анализ ч.1	110	**50,69**	92	**40,00**	98	**40,83**
Экономика	22	10,14	31	13,48	20	8,33
Английский язык	42	19,35	41	17,83	35	14,58
Дискретная математика	59	**27,19**	63	**27,39**	87	**36,25**
Программирование	57	**26,27**	18	7,83	46	19,17
Проектный семинар	40	18,43	43	18,70	34	14,17
Теоретические основы информатики	6	2,76	35	15,22	36	15,00
Философия	38	17,51	47	20,43	32	13,33

Жирным выделены значения, превышающие 25%.

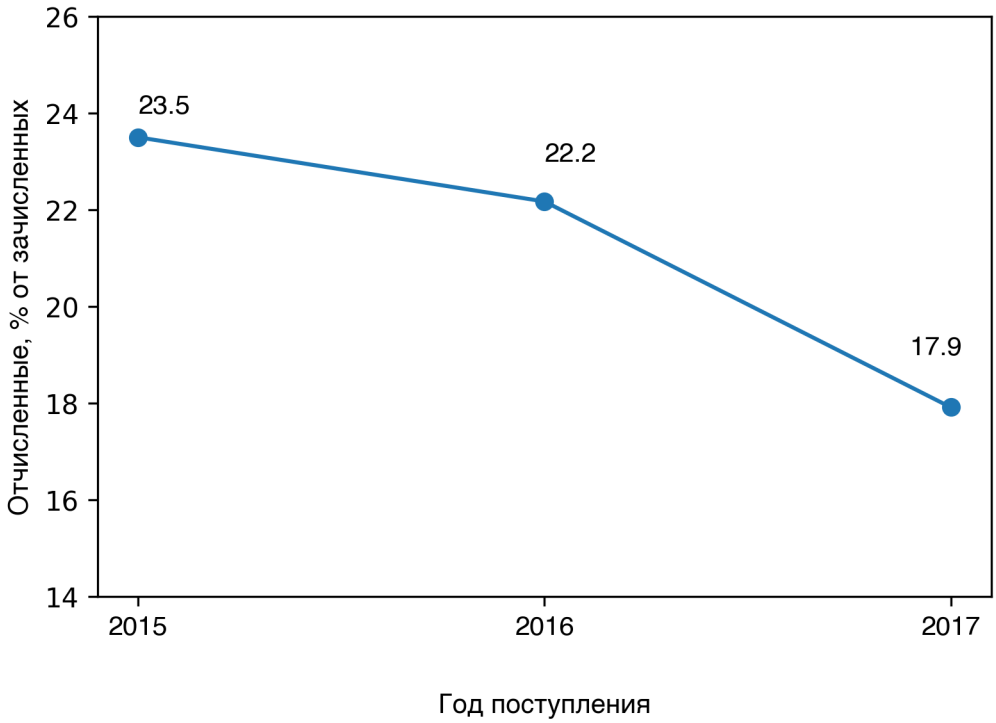

Рисунок 1. Процент отчислений в 2015–2017 гг.

Если брать во внимание конкретный момент отчисления (Рис. 2), больше всего отчислений приходится на второй семестр первого курса, то есть конец учебного года. Это сходится с результатами исследования [14] отчислений в НИУ ВШЭ в 2007–2009 гг. (пик на 1 модуле 2 курса соответствует периоду после пересдач дисциплин за 2 семестр 1 курса) и оправдывает использование такого периода наблюдений. Стоит отметить, что там же в качестве проблемных для факультета упоминаются дисциплины Линейная алгебра и Математический анализ. Отличие графика для 2016 от остальных, вероятно, связано с различиями в наполнении учебных планов. Так, например, только в 2016 году курс Математического анализа, который идёт два семестра подряд, начался не в первом семестре, а во втором.

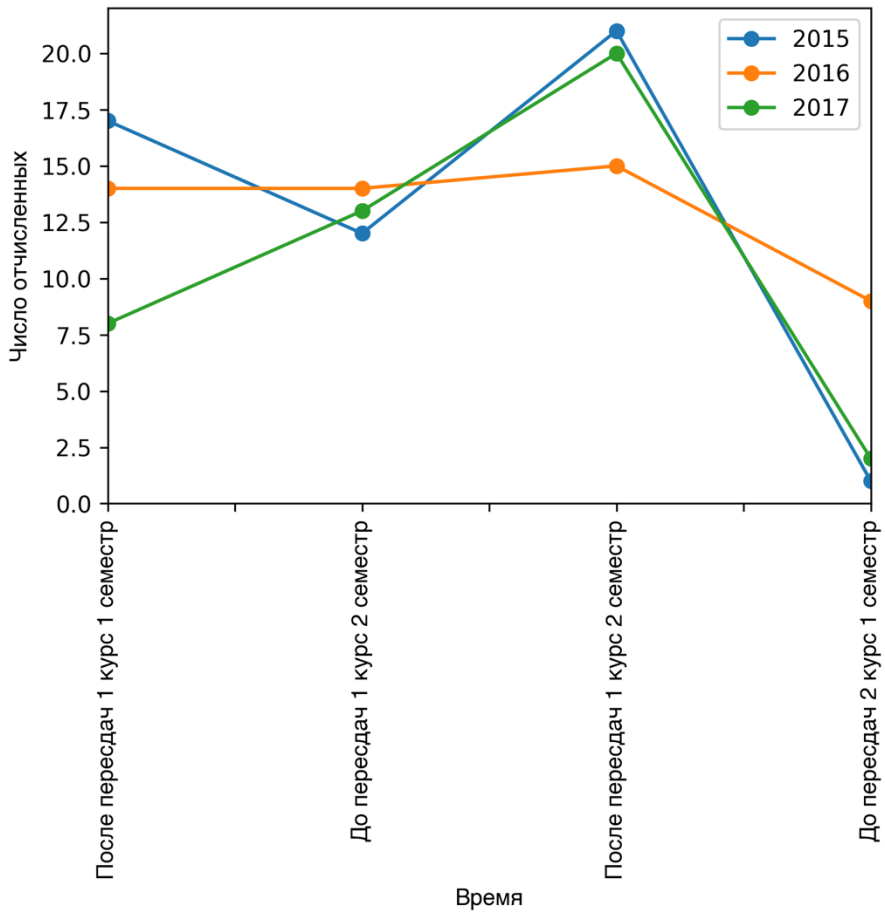

Рисунок 2. Число отчисленных по семестрам в 2015–2017 гг.

Рисунок 3. Из статьи [14]. Распределение функции риска быть отчисленным в зависимости от времени зачисления студента до наступления «академического тупика» (вместе рассматриваются данные когорт 2007–2009 гг.)

В ходе исследования выяснилось, что часть отчисленных поступает на программу вновь. Почти каждый год число повторно зачисленных студентов растет в абсолютном и процентном значениях. Вероятно, это связано с желанием студентов сохранить скидку (или, пересдав ЕГЭ, получить бюджетное место) или избежать оформления индивидуального учебного плана с повторным изучением неудовлетворительно сданных дисциплин — за эту возможность нужно доплатить, перейдя при этом на полную стоимость обучения. Но изучение причин повторных поступлений — тема для отдельного качественного исследования.

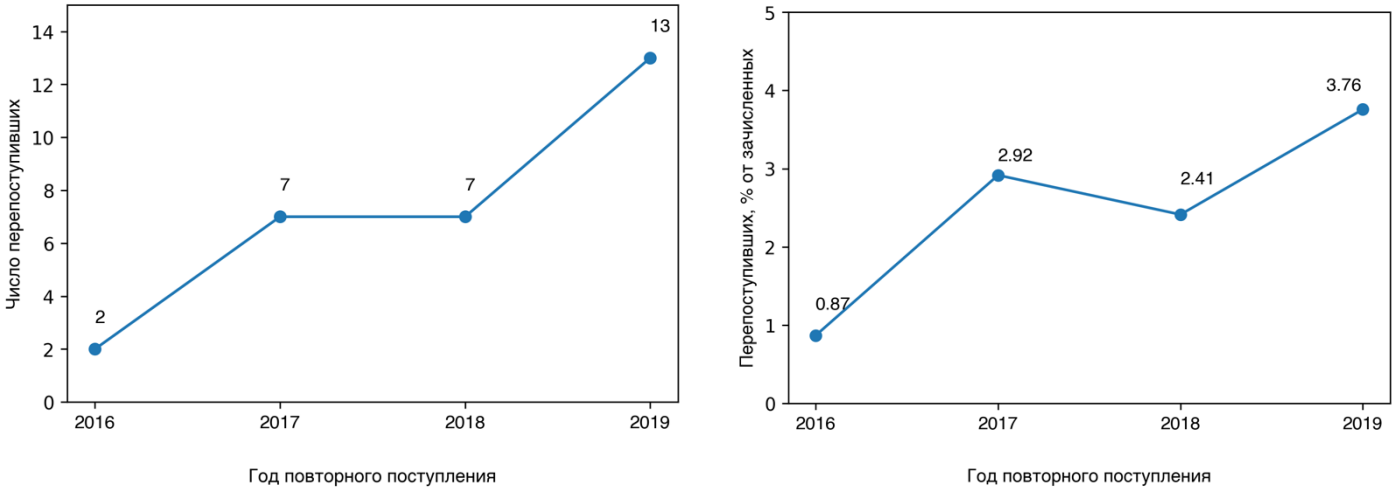

Рисунок 4. Число отчисленных студентов, которые поступили на программу заново в следующем году

4. МЕТОДОЛОГИЯ

Отчисление студента в контексте данного исследования представляет собой событие, которое может произойти по окончании первого или второго семестра первого курса. Для прогноза наступления этого события были созданы две бинарные переменные (*dropout_1* и *dropout_2* для отчисления после первого семестра или в конце года соответственно), принимающие значение 1, если событие произошло. Таким образом, задача прогноза отчислений была сведена к задаче получения бинарного классификатора с высокой предсказательной силой.

Сначала датасет был разделен на обучающую и тестовую выборки. Модели проверялись на 20% данных. Для использования в качестве зависимой переменной *dropout_2* из датасета были удалены отчисленные после первого семестра студенты. Далее были выбраны следующие модели:

1. Логистическая регрессия (*Logistic Regression*) — статистическая модель для прогнозирования вероятности наступления события, является классическим примером бинарного классификатора.
2. Дерево принятия решений (*Decision Tree*) — использование деревьев решений в этой работе обусловлено необходимостью в определении влияющих на отчисление факторов. Интерпретируемость отличает деревья от других применяемых в задачах классификации алгоритмов.
3. Метод опорных векторов (*Support Vector Machine*) — метод классификатора с максимальным зазором, разделяющий на классы по гиперплоскости; является примером обучения с учителем.

Классы наблюдений не сбалансированы, так как отчисленных в общей сложности в 8 раз меньше. Для оценки прогноза не использовалась метрика *accuracy* (доля правильно определённых классов для всей тестовой выборки) — цены ошибок неравнозначны, поэтому вместо нее для сравнения моделей были рассчитаны *precision* (1) и *recall* (2). При использовании вероятностей принадлежности к классу, а не самих классов в качестве вывода модели для классификаторов были построены ROC-кривые. Для определения оптимальных гиперпараметров моделей был использован *GridSearchCV*, выбирающий лучшую модель по F1 score (3) с помощью кросс-валидации. С полученными гиперпараметрами можно ознакомиться в Приложениях.

Таблица 5. Матрица сопряженности

		Прогноз	
		0	1
Реальность	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

$$precision(a, X) = \frac{TP}{TP + FP} \quad (1)$$

$$recall(a, X) = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \quad (3)$$

Наборы независимых переменных для разных классификаторов оставались неизменными, но для разных прогнозируемых переменных — *dropout_1* и *dropout_2* — они были разными. В качестве потенциальных предикторов академических исходов были выбраны следующие характеристики студентов:

Таблица 5. Используемые для прогноза переменные

y	$x_1, x_2 \dots x_n$
dropout_1	gender, enrol_year_2015, enrol_year_2016, enrol_year_2017, rank_1_1_before, avg_score_1_1_before, min_score_1_1_before, fails_1_1_before, rank_1_1_after, avg_score_1_1_after, min_score_1_1_after, fails_1_1_after, economics, lin_alg, safety (дисциплины)
dropout_2	gender, enrol_year_2015, enrol_year_2016, enrol_year_2017, rank_1_1_after, avg_score_1_1_after, min_score_1_1_after, fails_1_1_after, rank_1_2_after, avg_score_1_2_after, min_score_1_2_after, _1_2_after, rank_1_1_before, avg_score_1_1_before, min_score_1_1_before, fails_1_1_before, rank_1_2_before, avg_score_1_2_before, min_score_1_2_before, fails_1_2_before, eng, discrete_math, programming, lin_alg, calculus_1, economics, phil, history (дисциплины + *_before, _ after*)

На данном этапе можно предположить, что предсказание отчислений после первого семестра будет менее точным. Дело в том, что учебные планы одной и той же образовательной программы меняются из года в год, и общими для всех трёх рассматриваемых курсов в первом семестре были лишь Безопасность жизнедеятельности, Линейная алгебра и Экономика. Тем не менее, рейтинговая сумма и средний балл включают в себя оценки по всем остальным предметам.

5. РЕЗУЛЬТАТЫ

В идеальном случае модель прогноза должна работать так, что данных до пересдач будет достаточно для эффективного определения студентов в группе риска, однако для всех трёх использованных модификаций алгоритмов для первого семестра переменных с постфиксом «before» не хватает. Модели, в которых не задействованы результаты после пересдач, правильно предсказывают лишь половину отчислений и столько же или больше человек неправильно классифицируют как отчисленных. Включение в модели года поступления не повлияло на точность прогнозирования. Значимые улучшения говорили бы о том, что каждый год на отчисление влияют в разной мере разные факторы. Тем не менее, при переходе к моделям до пересдач после второго семестра, год поступления становится значимой переменной, повышающей точность модели.

Далее описываются результаты построения прогнозов с использованием всех доступных характеристик за первый семестр после пересдач и, в случае прогноза на конец года, с отсутствием и включением в модели данных после пересдач предметов второго семестра. Кривые ошибок классификаторов представлены на рисунках 9–11, показатели эффективности классификаторов — в таблице 6. Спецификации моделей находятся в Приложениях. Полученные модели не отличаются высокой точностью, когда целью является определение отчисленных студентов, а не продолживших обучение. Когда в спецификации оказывается слишком много переменных (модели после пересдач второго семестра), алгоритмы фокусируются только на оценках после пересдач, а точнее, на их отсутствии (когда студент, вероятно, уже просто забрал документы после трёх неудовлетворительных оценок) или просто рейтинге после. Так, например, были получены деревья на рисунке 8 с *recall* = 0.9 и 0.8 соответственно.

Таблица 6. Метрики

Модель		**Precision**	**Recall**	**AUC**
Logistic Regression	*dropout_1*	0,93	0,65	0,89
	dropout_2	0,83	1,00	0,99
	dropout_2 (after)	1,00	1,00	1,00
Decision Tree	*dropout_1*	1,00	0,65	0,85
	dropout_2	1,00	0,60	0,92
	dropout_2 (after)	0,83	1,00	0,98
Support Vector Machine	*dropout_1*	1,00	0,60	0,78
	dropout_2	0,86	0,60	0,98
	dropout_2 (after)	0,91	1,00	1,00

Дерево для *dropout_1* (*recall* = 0.65) опирается, в основном, на показатели после пересдач. Интересно, что среди всех построенных деревьев только в нём можно обнаружить переменную, отвечающую за пол студента. Дерево для *dropout_2* без результатов после пересдач за второй семестр обращает внимание на рейтинг первого семестра и Дискретную математику — одну из математических дисциплин с большим процентом пересдающих, если в модель не включены фиктивные переменные *enrol_year_2015, enrol_year_2016, enrol_year_2017*. Математические дисциплины есть и в дереве *dropout_2* с оценками после пересдач — к отчисленным причисляются студенты, не сдавшие Математический анализ, Дискретную математику и Программирование.

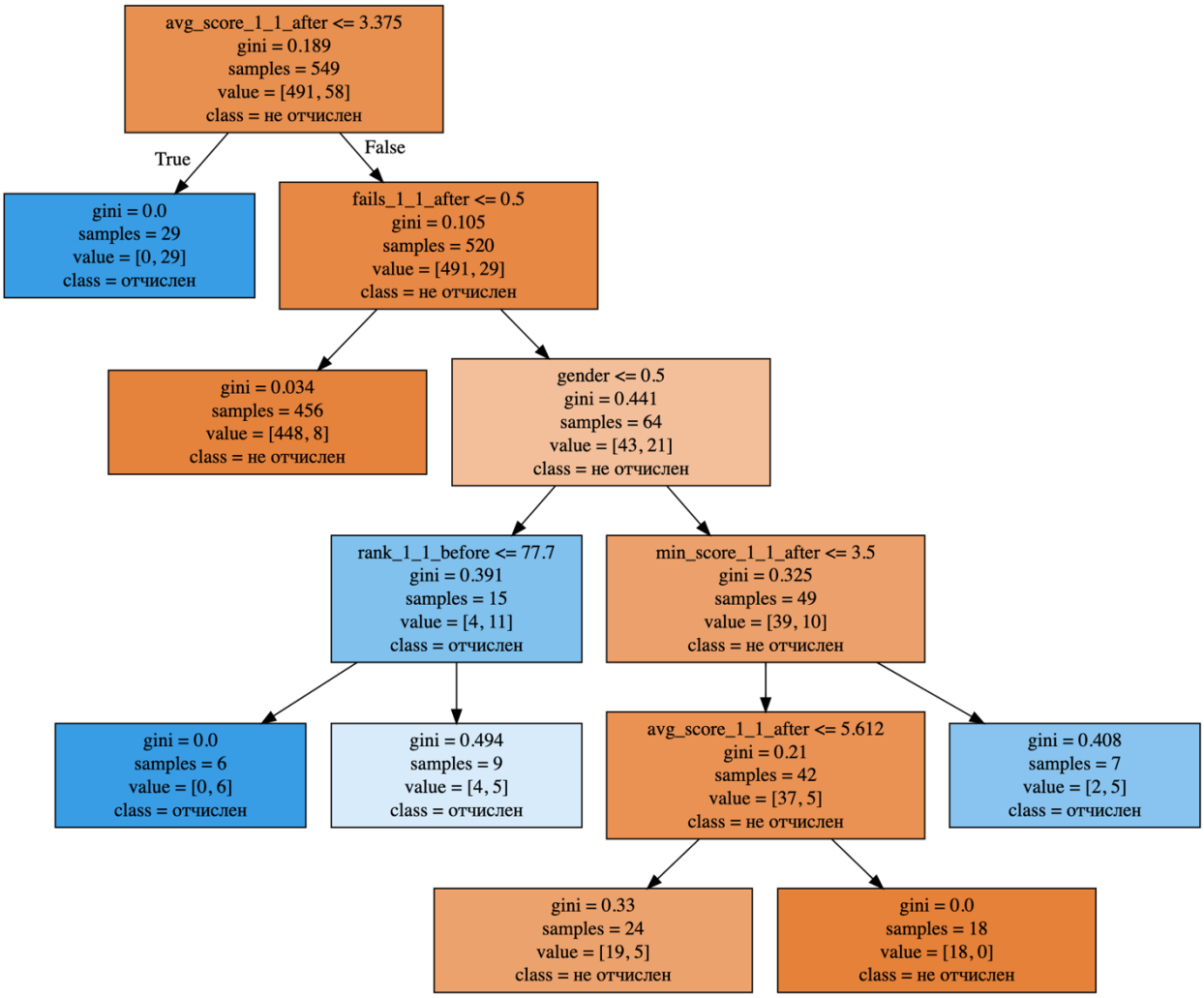

Рисунок 5. Дерево решений для *dropout_1*

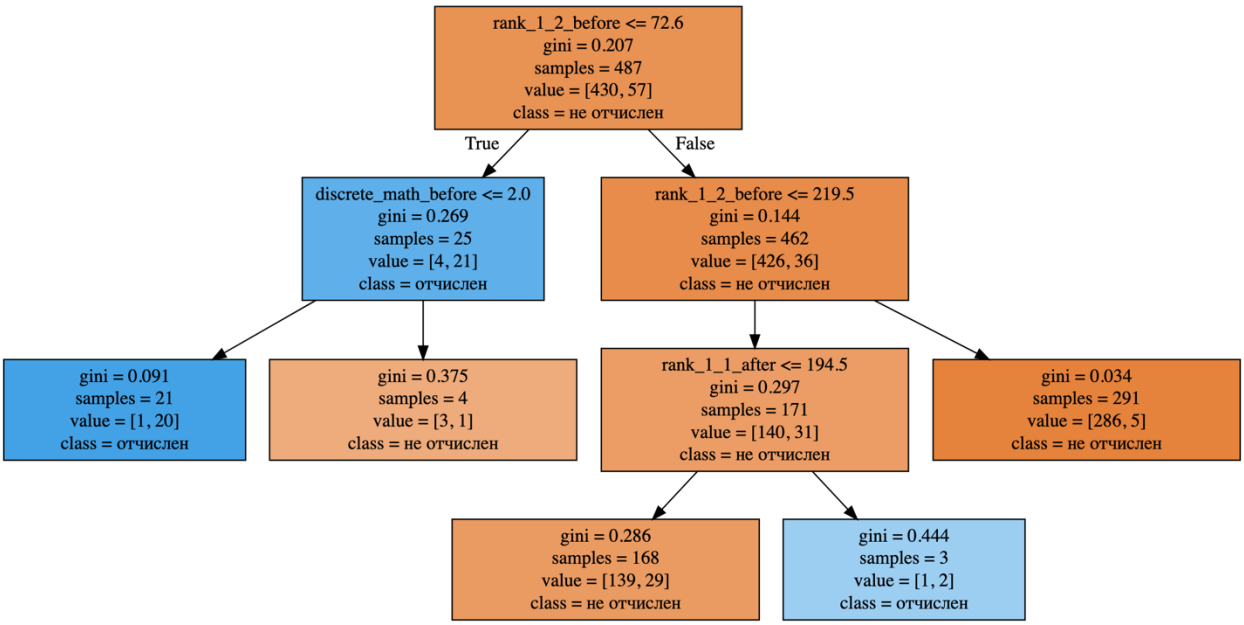

Рисунок 6. Дерево решений для *dropout_2* до пересдач без года поступления

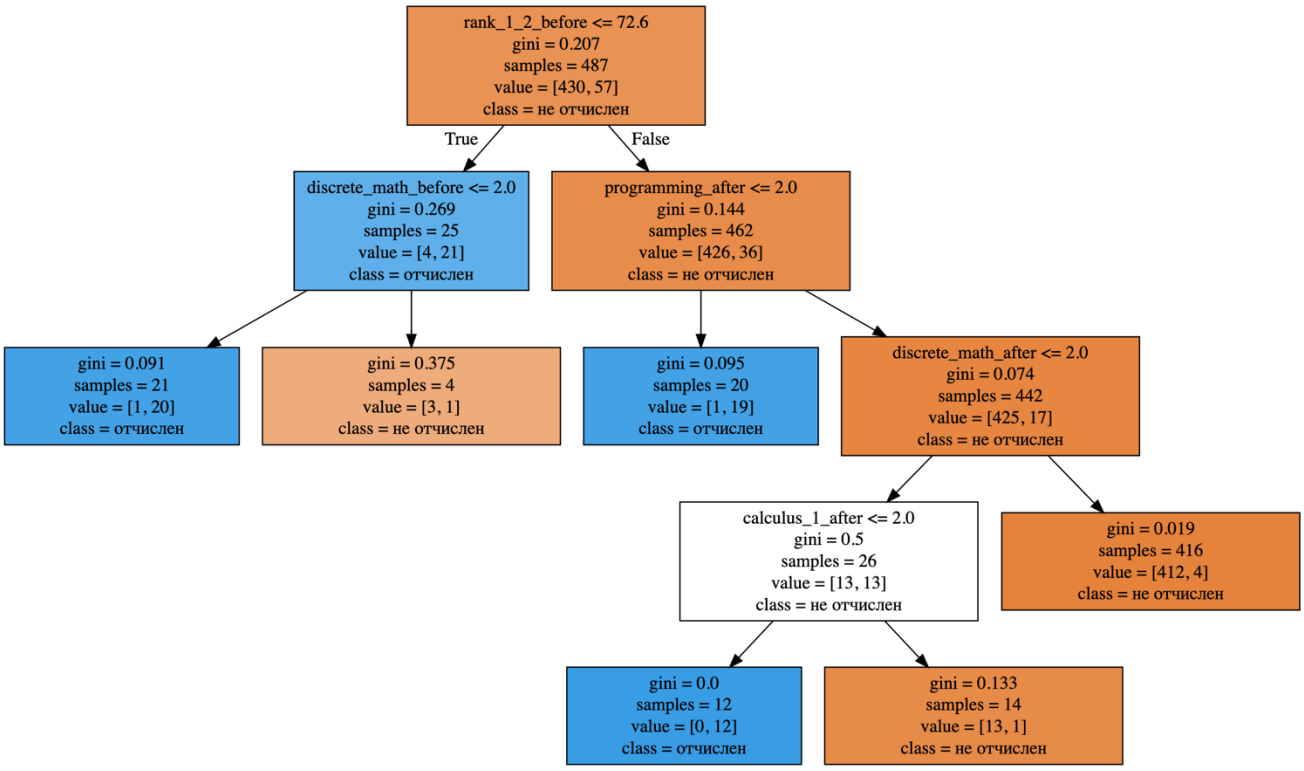

Рисунок 7. Дерево решений для *dropout_2* после пересдач

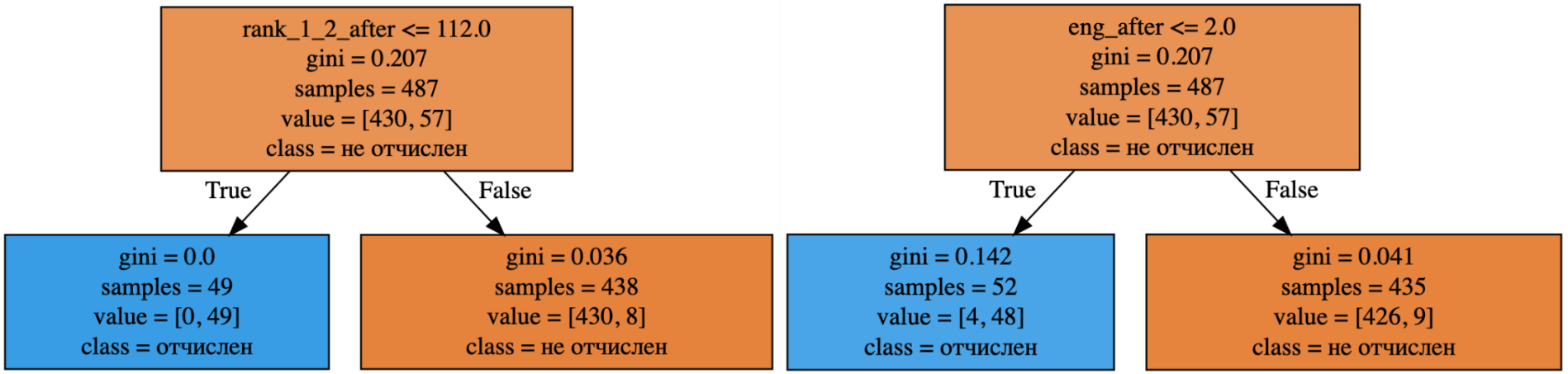

Рисунок 8. Некоторые из деревьев для *dropout_2* после пересдач

С учётом года зачисления модели *dropout_2* до пересдач имеют *precision* и *recall* > 0.9, а логистическая регрессия безошибочно определяет все наблюдения в соответствующие классы. Сопоставимой точностью обладают модели для *dropout_2* с включением оценок после пересдач (*recall* = 1 для всех алгоритмов), так как на вход подаётся достаточно исчерпывающая информация о уже прошедшем семестре и осеннем периоде пересдач.

На рисунках ниже представлены графики, позволяющие сравнить модели между собой в общем. ROC-кривая показывает производительность моделей на всех возможных классификационных порогах. Стоит отметить, что ROC-кривые в данном случае как показатель эффективности должны быть использованы с осторожностью, так как для несбалансированных классов это может привести к ошибочным суждениям. Диагональ (на рисунках она красным пунктиром) представляет из себя бесполезный классификатор, для идеального же классификатора кривая должна проходить через левый верхний угол. Таким образом, наиболее близкая к диагонали кривая говорит о низкой эффективности метода. На практике близость кривых определяется оценкой площади под ними — AUC (*area under the ROC curve*) — так как не всегда по рисунку можно сказать, кривая какого классификатора на самом деле выше. Значение AUC = 0.5 соответствует диагонали и говорит о равной вероятности определения в тот или иной класс. Если смотреть на три графика одновременно, можно заметить, что чем больше данных вошло в модель, тем ближе к левому верхнему углу оказались кривые, а логистическая регрессия во всех случаях имеет наибольшее значение AUC (как и *precision* и *recall*).

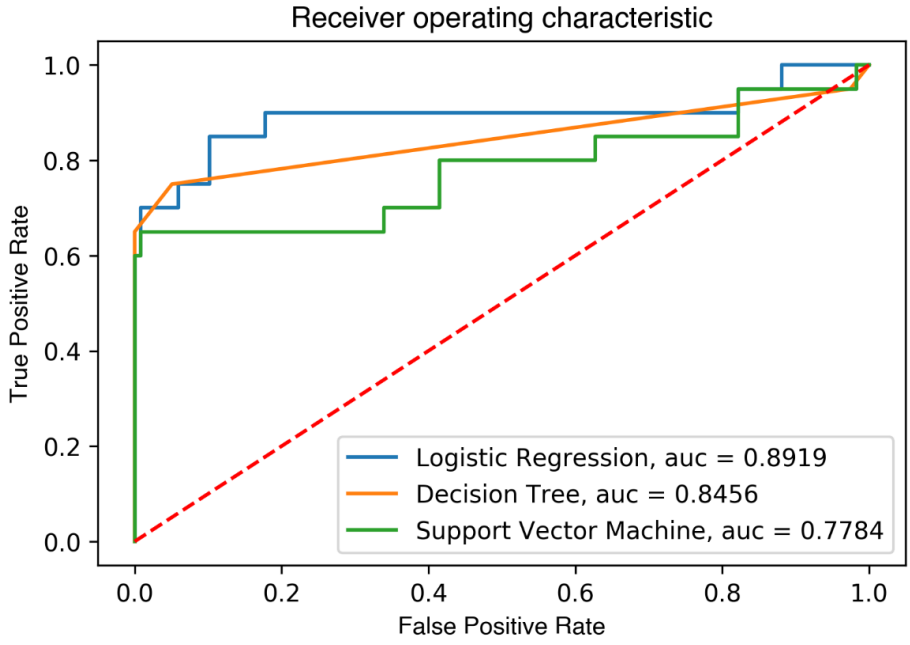

Рисунок 9. ROC-кривые для *dropout_1*

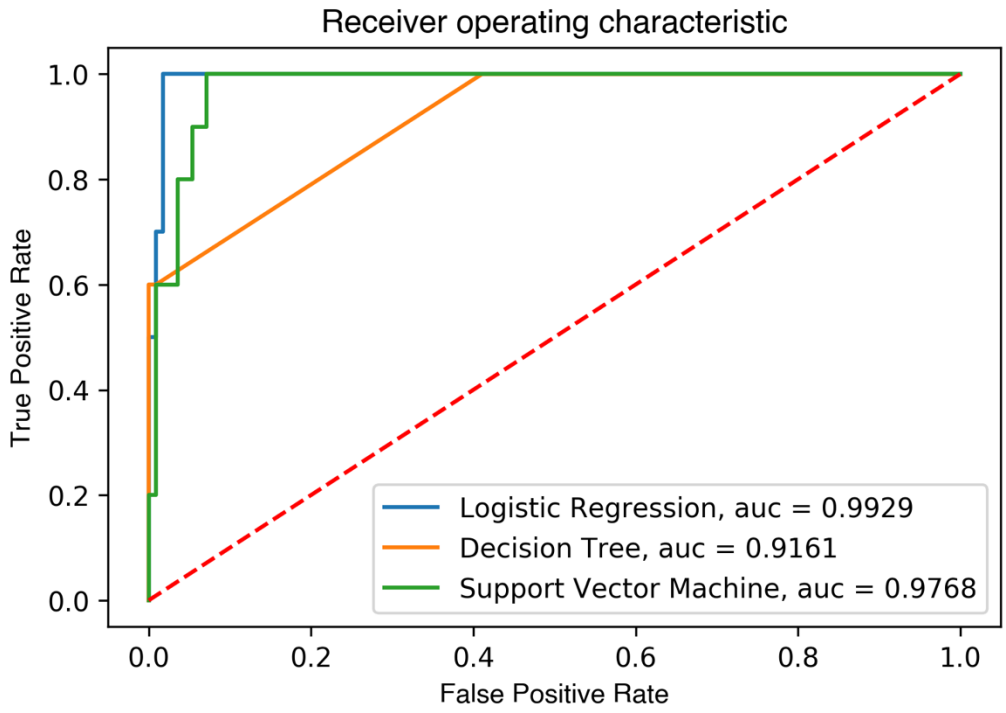

Рисунок 10. ROC-кривые для *dropout_2* без года поступления

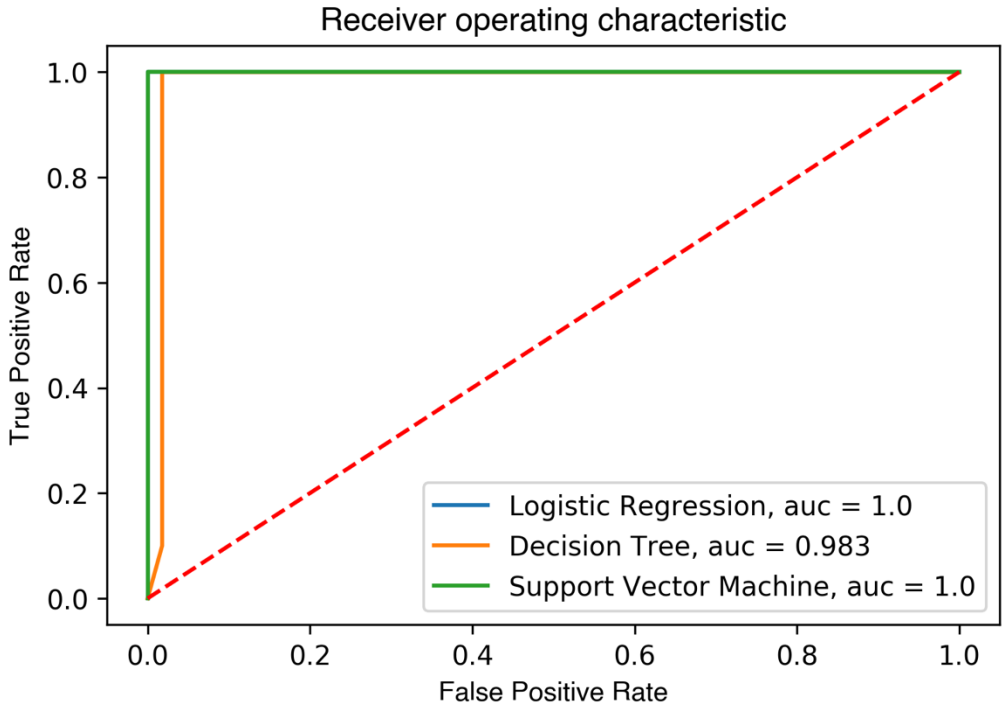

Рисунок 12. ROC-кривые для *dropout_2* после пересдач

6. ЗАКЛЮЧЕНИЕ

В данной работе был проведен обзор литературы, посвященной проблеме студенческого отсева, и, на основе существующей теоретической рамки, анализ рейтингов студентов ОП «Бизнес-информатика» НИУ ВШЭ за 2015–2019 годы. В процессе анализа были описаны когорты первокурсников 2015–2017 годов поступления, оценены масштабы отчислений на разных этапах обучения, выделены проблемные дисциплины (ими оказались математические курсы). Результатом проведённого исследования стали модели бинарной классификации, позволяющие с высокой точностью определить отчисляемых студентов, но лишь на этапе, следующим за проведением пересдач. Так, были построены модели логистической регрессии, деревья решений, так же для задачи классификации был задействован метод опорных векторов. При прочих равных логистическая регрессия показывала лучший результат. Тем не менее, на данных до пересдач в первом семестре доля угадываемых алгоритмами отчисленных не превысила 65%. Это может быть последствием ограничений, накладываемых доступными данными, их неполнотой, а также различиями в учебной нагрузке между потоками разных лет. Возможно, включение в модели результатов промежуточной аттестации студентов позволило бы выделить определенные факторы, которые влияют на отсев ещё в предэкзаменационный период. Так, подтверждается предположение о том, что чем дальше во времени уходят данные, тем точнее можно сделать прогноз, особенно учитывая, что большее число отчислений приходится именно на конец первого курса.

Дальнейшая работа над схожими данными может выявить некоторые упущенные в этом исследовании зависимости, поскольку даже выборка из всего трех первых курсов дает представление о процессе обучения на ОП и возможных факторах выбытия. Информация о студентах, использованная в этой работе, может быть дополнена данными о текущей успеваемости (возможно, это позволит описать зависимость между оценками за промежуточный контроль и экзамен) и записями на пересдачи. Подробнее могут быть рассмотрены кейсы повторно зачисленных на факультет и повторно изучающих курсы по специальному учебному плану студентов: становятся ли изменения в стандартной образовательной траектории толчком для изменений в успеваемости? Анализ можно распространить и на другие образовательные программы университета.

7. ИСТОЧНИКИ

1. Рейтинги студентов бакалавриата образовательной программы "Бизнес-информатика" [Электронный ресурс]: https://www.hse.ru/ba/bi/ratings/archive (дата обращения 01.05.2020)
2. Груздев И.А., Горбунова Е.В., Фрумин И.Д. Студенческий отсев в российских вузах: к постановке проблемы // Вопросы образования. 2013. №2. С. 67–81
3. Education at a Glance 2010: OECD Indicators https://doi.org/10.1787/eag-2010-en
4. Frobisher A. Degree Dropouts [Электронный ресурс]: https://debut.careers/insight/degree-dropouts/ (дата обращения 01.05.2020)
5. Кочергина Е. В., Прахов И.А. Взаимосвязь между отношением к риску, успеваемостью студентов и вероятностью отчисления из вуза // Вопросы образования 2016. № 4. С. 206–228
6. Хавенсон Т.Е., Соловьева А.А. Связь результатов Единого государственного экзамена и успеваемости в вузе // Вопросы образования 2014. № 1. С. 176–199
7. Горбунова Е.В., Кондратьева О.С. Анализ гендерных различий в выбытии из вуза российских и американских студентов программ бакалавриата // Universitas 2013. № 3 (Том 1). С. 48–69
8. Донец Е. Опыт исследования студенческих отчислений на примере МГУ// Мониторинг университета. 2011. № 6. С. 33–38
9. Замков О.О., Пересецкий А.А. Динамика влияния оценок ЕГЭ на последующие результаты студентов бакалавриата МИЭФ НИУ ВШЭ // XIV Апрельская международная научная конференция по проблемам развития экономики и общества, 2013. М.: Издательский дом НИУ ВШЭ, 2014. Книга 4 С. 40–49
10. Замков О.О. Эконометрический анализ факторов академических достижений студентов в МИЭФ НИУ ВШЭ // XII Международная научная конференция по проблемам развития экономики и общества, 2011. М.: Издательский дом НИУ ВШЭ, 2012. Книга 2 С. 384-391
11. Рафаилов Р. Л. «Эконометрический анализ академических успехов студентов МИЭФ и Peer effects». Международный институт экономики и финансов, 2015
12. Alban, Mayra, & David Mauricio. (2019). Predicting University Dropout through Data Mining: A Systematic Literature. Indian Journal of Science and Technology, 12

13. Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2018). Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods
14. Горбунова, Е. В. Изучение отчислений студентов в бакалавриате/специалитете НИУ ВШЭ // Мониторинг университета. 2011. № 6. С. 22–32

8. ПРИЛОЖЕНИЯ

Приложение 1. Спецификации моделей

		Hyperparameters	**Features**
Logistic Regression	*dropout_1*	C: 1.0, class_weight: None, l1_ratio: None, max_iter: 100, multi_class: auto, n_jobs: None, penalty: l2, random_state: 333, solver: newton-cg, tol: 0.0001	rank_1_1_after, avg_score_1_1_after, min_score_1_1_after, fails_1_1_after, rank_1_1_before, avg_score_1_1_before, min_score_1_1_before, fails_1_1_before, gender, economics_before, lin_alg_before
	dropout_2	C: 1.0, class_weight: None, l1_ratio: None, max_iter: 100, multi_class: auto, n_jobs: None, penalty: l2, random_state: 333, solver: liblinear, tol: 0.0001	gender, rank_1_2_before, fails_1_2_before, rank_1_1_before, avg_score_1_1_before, min_score_1_1_before, fails_1_1_before, rank_1_1_after, avg_score_1_1_after, min_score_1_1_after, fails_1_1_after, eng_before, discrete_math_before, programming_before, lin_alg_before, calculus_1_before, economics_before
	dropout_2 (after)	C: 1.0, class_weight: None, l1_ratio: None, max_iter: 100, multi_class: auto, n_jobs: None, penalty: l2, random_state: 333, solver: newton-cg, tol: 0.0001	gender, rank_1_2_before, fails_1_2_before, eng_before, discrete_math_before, programming_before, lin_alg_before, calculus_1_before, economics_before, eng_after, discrete_math_after, programming_after, lin_alg_after, calculus_1_after, economics_after
Decision Tree	*dropout_1*	ccp_alpha: 0.0, class_weight: None, criterion: gini, max_depth: None, max_features: None, max_leaf_nodes: 7, min_impurity_decrease: 0.0, min_impurity_split: None, min_samples_leaf: 1, min_samples_split: 10, min_weight_fraction_leaf: 0.0, random_state: 333, splitter: best	rank_1_1_after, avg_score_1_1_after, min_score_1_1_after, fails_1_1_after, rank_1_1_before, avg_score_1_1_before, min_score_1_1_before, fails_1_1_before, gender, economics_before, lin_alg_before

	dropout_2	ccp_alpha: 0.0, class_weight: None, criterion: gini, max_depth: None, max_features: log2, max_leaf_nodes: 5, min_impurity_decrease: 0.0, min_impurity_split: None, min_samples_leaf: 1, min_samples_split: 20, min_weight_fraction_leaf: 0.0, random_state: 333, splitter: best	gender, rank_1_2_before, fails_1_2_before, rank_1_1_before, avg_score_1_1_before, min_score_1_1_before, fails_1_1_before, rank_1_1_after, avg_score_1_1_after, min_score_1_1_after, fails_1_1_after, eng_before, discrete_math_before, programming_before, lin_alg_before, calculus_1_before, economics_before
	dropout_2 (after)	ccp_alpha: 0.0, class_weight: None, criterion: gini, max_depth: None, max_features: 5, max_leaf_nodes: 6, min_impurity_decrease: 0.0, min_impurity_split: None, min_samples_leaf: 1, min_samples_split: 10, min_weight_fraction_leaf: 0.0, random_state: 333, splitter: best	gender, rank_1_2_before, fails_1_2_before, eng_before, discrete_math_before, programming_before, lin_alg_before, calculus_1_before, economics_before, eng_after, discrete_math_after, programming_after, lin_alg_after, calculus_1_after, economics_after
Support Vector Machine	*dropout_1*	C: 1.0, break_ties: False, cache_size: 200, class_weight: 0: 8, 1: 1}, coef0: 0.0, decision_function_shape: ovr, degree: 3, gamma: scale, kernel: rbf, probability: False, random_state: 333, shrinking: True	rank_1_1_after, avg_score_1_1_after, min_score_1_1_after, fails_1_1_after, rank_1_1_before, avg_score_1_1_before, min_score_1_1_before, fails_1_1_before, gender, economics_before, lin_alg_before
	dropout_2	C: 1.0, break_ties: False, cache_size: 200, class_weight: None, coef0: 0.0, decision_function_shape: ovr, degree: 3, gamma: scale, kernel: rbf, probability: False, random_state: 333, shrinking: True	gender, rank_1_2_before, fails_1_2_before, rank_1_1_before, avg_score_1_1_before, min_score_1_1_before, fails_1_1_before, rank_1_1_after, avg_score_1_1_after, min_score_1_1_after, fails_1_1_after, eng_before, discrete_math_before, programming_before, lin_alg_before, calculus_1_before, economics_before
	dropout_2 (after)	C: 1.0, break_ties: False, cache_size: 200, class_weight: None, coef0: 0.0, decision_function_shape: ovr, degree: 3, gamma: scale, kernel: linear, probability: False, random_state: 333, shrinking: True	gender, rank_1_2_before, fails_1_2_before, eng_before, discrete_math_before, programming_before, lin_alg_before, calculus_1_before, economics_before, eng_after, discrete_math_after, programming_after, lin_alg_after, calculus_1_after, economics_after