

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет

вычислительной техники

Кафедра

«Вычислительная техника»

Направление подготовки: 09.04.01 «Информатика и вычислительная техника»

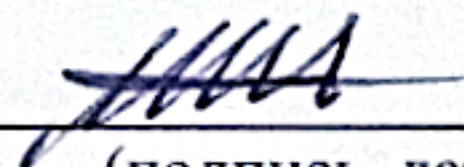
Магистерская программа: «Прикладной искусственный интеллект»

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

на тему

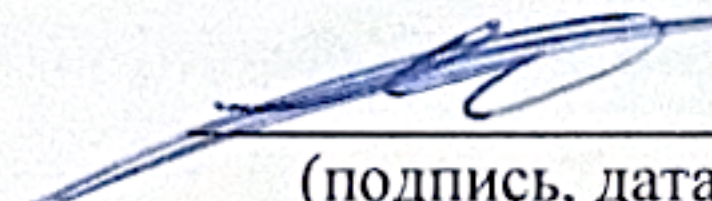
СИСТЕМА ПРОГНОЗИРОВАНИЯ
АКАДЕМИЧЕСКОЙ ЗАДОЛЖЕННОСТИ НА
ОСНОВЕ ТЕКУЩЕЙ УСПЕВАЕМОСТИ И
ПРОЦЕНТА ПРОПУСКОВ

Студент


(подпись, дата) 26.05.2026

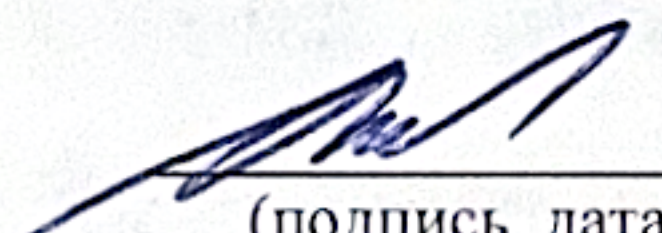
Мальков Илья Андреевич
(ФИО полностью)

Руководитель


(подпись, дата) 26.05.2026

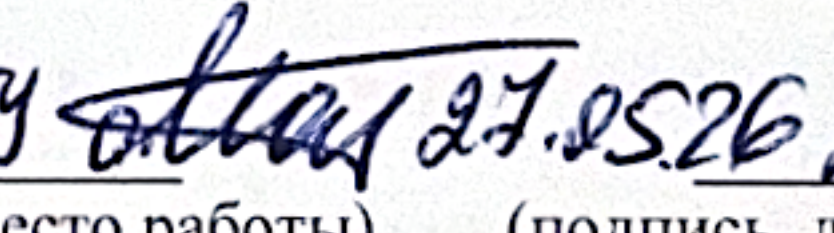
Синев М.П.
(фамилия, инициалы)

Нормоконтролёр


(подпись, дата) 26.05.2026

Бычков А.С.
(фамилия, инициалы)

Рецензент

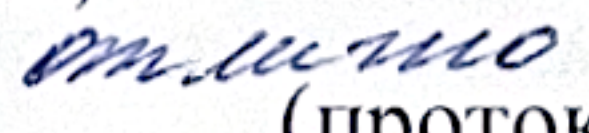
зав. каф. Программирования ИИ ПГУ  27.05.26 Мартынов А.И.
(должность, место работы) (подпись, дата)
(фамилия, инициалы)

Работа допущена к защите (протокол зас. кафедры от 22.05.2026 № 14)

Заведующий кафедрой

(подпись)

М.А. Митрохин
(фамилия, инициалы)

Работа защищена с отметкой 

(протокол заседания ГЭК от 04.06.2026 № 7)

Секретарь ГЭК

(подпись)

Е.И. Гурин
(фамилия, инициалы)

Пенза 2026

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет

вычислительной техники

Кафедра

«Вычислительная техника»

«Утверждаю»

Заведующий кафедрой

М.А. Митрохин

« 20 » 09 20 24 г.

**ЗАДАНИЕ
НА МАГИСТЕРСКУЮ ДИССЕРТАЦИЮ**

1. Студент Мальков Илья Андреевич гр.24ВВИм1 факультета ВТ
2. Тема работы «Система прогнозирования академической задолженности на основе текущей успеваемости и процента пропусков»

Тема утверждена приказом ПГУ № 920/0 от "20" сентября 2024г.

3. Руководитель магистерской диссертации к.т.н., доцент Синев М.П.
4. Задание на работу (назначение разработки, исходные данные и т.п.)

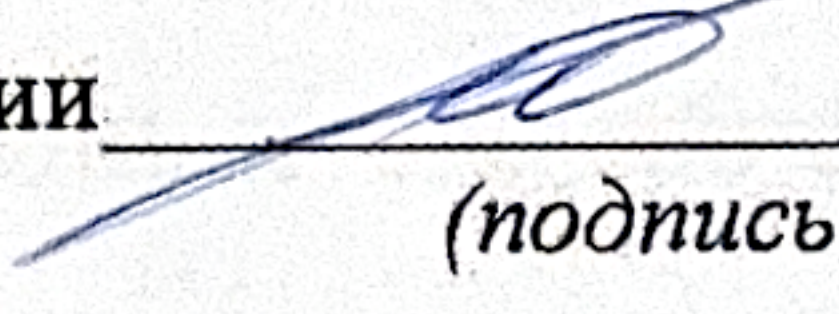
Назначение разработки: разработка системы прогнозирования академической задолженности на основе текущей успеваемости и процента пропусков.

Разрабатываемая система должна выполнять автоматический прогноз наличия академической задолженности у учащихся на основе анализа текущей успеваемости и процента пропущенных уроков.

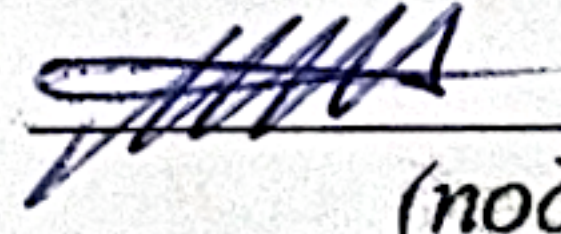
5. Перечень подлежащих разработке вопросов
 1. Анализ предметной области.
 2. Изучение литературы по теме исследования.
 3. Разработка алгоритма прогнозирования академической задолженности.
 4. Тестирование разработанной системы, анализ полученных результатов.
 5. Подтверждение гипотезы исследования.
 6. Календарный график выполнения работы

№ п/п	Наименование этапов работы	Объем работы	Срок выполнения	Подпись руководителя
1.	Изучение предметной области	11%	20.09.2024- 29.01.2025	
2.	Анализ требований	11%	29.01.2025- 30.03.2025	
3.	Изучение литературы по теме исследования	22%	30.03.2025- 30.05.2025	
4.	Проектирование и реализация системы	27%	01.06.2025- 01.09.2025	
5.	Тестирование	17%	02.09.2025- 02.10.2025	
6.	Анализ полученных результатов исследования	6%	01.04.2026- 29.04.2026	
7.	Оформление пояснительной записки	6%	30.04.2026- 20.05.2026	

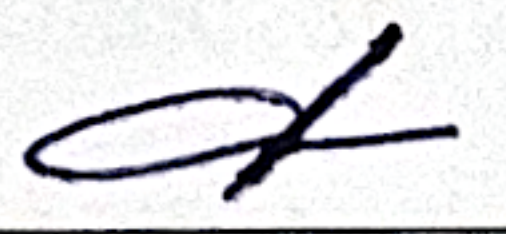
Дата выдачи задания " 20 " сентября 20 24 г.

Руководитель магистерской диссертации 
(подпись, дата)

Синев М.П.

Задание к исполнению принял студент 
(подпись, дата)

Мальков И.А.

Работу к защите допустить
Декан факультета 
(подпись, дата)

Фионова Л.Р.

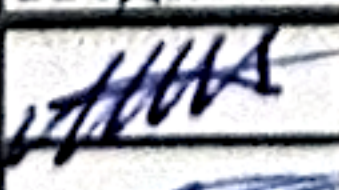
01.06.2026

РЕФЕРАТ

Пояснительная записка содержит 83 листа, 11 рисунков, 8 таблиц, 12 источников, 2 приложения.

АКАДЕМИЧЕСКАЯ ЗАДОЛЖЕННОСТЬ, ПРОГНОЗИРОВАНИЕ УСПЕВАЕМОСТИ, БИНАРНАЯ КЛАССИФИКАЦИЯ, ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ, СЛУЧАЙНЫЙ ЛЕС, РИСК-СКОР, ПОСЕЩАЕМОСТЬ, ЭЛЕКТРОННЫЙ ЖУРНАЛ, PYTHON, SCIKIT-LEARN, GOOGLE COLAB

Целью работы является построение и проверка прогностической модели, способной заблаговременно выявлять учащихся с высоким риском академической задолженности на основе минимального набора данных, типичного для школьного электронного журнала, таких как текущая успеваемость и пропуска. В отличие от большинства решений, нацеленных на высшее образование и использующих труднодоступные признаки, работа показывает достаточность лишь тех показателей, которые уже присутствуют в любом электронном журнале. В ходе исследования сформирован прототип, включающий загрузку открытого датасета Student Performance, конструирование признака риск-скор на основе корреляционного анализа, обучение логистической регрессии и случайного леса, оценку качества и интерпретацию метрик. Проект выполнен на Python в среде Google Colab (pandas, scikit-learn, matplotlib). Логистическая регрессия достигла точности 87% и полноты выявления задолженностей 79%; переход к случайному лесу не дал статистически значимого улучшения, а линейная модель сохранила преимущество в интерпретируемости. Проведённый анализ подтверждает, что предложенный подход может служить методической основой для создания систем раннего оповещения в общеобразовательных учреждениях.

					ПГУ 1.090401.14.001 ПЗ				
Изм.	Лист	№ докум	Подпись	Дата	Система прогнозирования академической задолженности на основе текущей успеваемости и процента пропусков Пояснительная записка	Лит.		Лист	Листов
Разраб.		Мальков И.А.		26.05.26				2	83
Провер.		Синев М.П.		26.05.26					
Н. Контр.		Бычков А.С.		26.05.26					
Утв.		Митрохин М.А.		26.05.26					
						ФВТ, каф. ВТ гр. 24ВВим1			

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	9
1 Обзор литературы и существующих решений прогнозирования	13
1.1 Факторы, влияющие на академическую задолженность	13
1.2 Методы машинного обучения в прогнозировании успеваемости	15
1.3 Существующие системы раннего предупреждения	16
1.4 Выводы по обзору литературы	20
2 Материалы и методы прогнозирования	23
2.1 Описание датасета.....	23
2.2 Подготовка данных и конструирование признаков	26
2.3 Методы моделирования.....	28
2.4 Процедура обучения и оценки.....	30
2.5 Инструментарий	33
2.6 Выводы по главе 2.....	34
3 Анализ результатов исследования.....	36
3.1 Разведочный анализ данных	36
3.2 Результаты моделирования логистической регрессии.....	42
3.3 Сравнение с моделью случайного леса	45
3.4 Интерпретация баланса пропуска должника и ложной тревогой	48
3.5 Сравнение с результатами других исследований	51
3.6 Подтверждение гипотезы	53
3.7 Ограничения исследования.....	55
3.8 Практическая значимость и направления развития	57
3.9 Выводы по экспериментальной части	60
ЗАКЛЮЧЕНИЕ	62
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	66

ПРИЛОЖЕНИЕ А	68
ПРИЛОЖЕНИЕ Б.....	74

ВВЕДЕНИЕ

Современные образовательные учреждения всё чаще сталкиваются с проблемой академической задолженности учащихся, которая не только снижает индивидуальные показатели успеваемости, но и увеличивает риск отчисления, ухудшает репутацию организации и требует дополнительных ресурсов для ликвидации последствий. Традиционные методы контроля, основанные на ручном анализе журналов и эпизодическом мониторинге со стороны преподавателей, оказываются недостаточно эффективными в условиях большого объёма накапливаемых данных. В этой связи особое значение приобретают технологии предсказания, позволяющие заблаговременно выявлять учащихся, находящихся в зоне риска, и принимать меры до формального возникновения задолженности. Многочисленные исследования в области Educational Data Mining (EDM) подтверждают, что текущая успеваемость и уровень посещаемости являются одними из наиболее значимых факторов, влияющих на будущие академические трудности, однако практическое внедрение прогнозных моделей в российских учебных заведениях остаётся ограниченным.

В зарубежной литературе накоплен значительный опыт применения методов машинного обучения для предсказания учебной неуспеваемости: используются логистическая регрессия, деревья решений, случайный лес, градиентный бустинг и другие алгоритмы. Работы, выполненные на открытых наборах данных (например, Student Performance Data Set), демонстрируют принципиальную возможность построения моделей с приемлемыми показателями точности и полноты. Вместе с тем многие из этих решений ориентированы на вузовскую среду, иностранные системы оценивания либо требуют большого числа разнородных признаков, не всегда доступных в условиях средней школы. Таким образом, существует потребность в исследовании, которое на основе ограниченного, но типичного для школьного электронного журнала набора показателей позволило бы построить

интерпретируемую модель, пригодную для дальнейшей адаптации к российским реалиям.

Цель исследования – выявить характер и силу связи между текущей успеваемостью, пропусками занятий и риском академической задолженности, а также построить на этой основе прогностическую модель, способную с приемлемой точностью классифицировать учащихся по степени риска, используя открытые данные о школьной успеваемости.

Объект исследования – образовательные данные учащихся, представленные в виде структурированных записей об оценках, посещаемости и сопутствующих характеристиках.

Предмет исследования – взаимосвязь между показателями текущей успеваемости, пропусков и вероятностью возникновения академической задолженности, а также методы машинного обучения, позволяющие формализовать эту взаимосвязь.

Гипотеза исследования заключается в том, что текущие оценки и количество пропусков являются статистически значимыми предикторами итоговой академической задолженности, а также построенная на их основе модель бинарной классификации способна идентифицировать учащихся группы риска с полнотой, достаточной для практического использования.

Для достижения поставленной цели и проверки гипотезы были определены следующие задачи:

- провести анализ факторов, влияющих на возникновение академической задолженности, по данным научной литературы и существующих систем-аналогов,

- выполнить разведочный анализ открытого датасета «Student Performance», оценить распределения ключевых переменных и корреляционные связи между ними,

- на основе исходных переменных сконструировать информативные признаки (средний балл за период, композитный риск-скор), теоретически и эмпирически обосновать их состав,

- построить модель логистической регрессии с выбранным набором признаков и дополнительно обучить модель случайного леса для сопоставления результатов,

- провести серию экспериментов, таких как варьирование состава признаков, изменение порога бинаризации целевой переменной и порога классификации, оценить влияние этих модификаций на основные метрики качества,

- на основе полученных результатов подтвердить или опровергнуть выдвинутую гипотезу, сформулировать практические рекомендации и направления дальнейших работ.

Исследование носит эмпирико-аналитический характер и опирается на методы разведочного анализа данных, корреляционный анализ, бинаризацию целевой переменной, конструирование признаков на основе предметных знаний, а также методы машинного обучения – логистическую регрессию и случайный лес. Оценка качества моделей проводилась с использованием метрик accuracy, precision, recall, F1-меры и матрицы ошибок. Все вычислительные эксперименты выполнены на языке Python 3 в среде Google Colab с применением библиотек pandas, numpy, matplotlib и scikit-learn. Полный код прототипа представлен в Приложении.

Практическая значимость работы заключается в том, что полученные результаты могут служить методической и алгоритмической основой для создания системы раннего оповещения о риске академической задолженности в общеобразовательных учреждениях. Выявленные закономерности и построенная модель демонстрируют возможность эффективного использования ограниченного набора данных, типичного для электронных журналов, и могут быть адаптированы к условиям российской школы после валидации на реальных данных.

Научная новизна исследования состоит в его точечной ориентации на условия, характерные для общеобразовательной школы. В отличие от большинства работ, сосредоточенных на высших учебных заведениях

или использующих широкий круг труднодоступных признаков (психологических, социальных, поведенческих), данная работа демонстрирует, что для уверенного прогнозирования академической задолженности достаточно минимального набора показателей – текущих оценок и пропусков, – который присутствует в любом типовом электронном журнале. Такой подход приближает методы Educational Data Mining к повседневной школьной практике и обосновывает возможность построения простых, интерпретируемых моделей, пригодных для использования педагогом без специальной технической подготовки.

1 Обзор литературы и существующих решений прогнозирования

1.1 Факторы, влияющие на академическую задолженность

Академическая задолженность представляет собой формальное состояние, при котором учащийся не выполнил требования образовательной программы по одному или нескольким предметам в установленные сроки. В нормативном смысле она возникает в результате неудовлетворительного прохождения промежуточной аттестации, отсутствия на аттестационных мероприятиях без уважительной причины либо недопуска к ним вследствие систематических пропусков занятий [1]. Задолженность запускает цепочку негативных последствий: для учащегося – дополнительную учебную нагрузку, психологический стресс, риск отчисления, а для образовательной организации – рост показателей неуспешности, необходимость выделения ресурсов на ликвидационные процедуры, потенциальное снижение репутации [2].

Многочисленные эмпирические исследования, выполненные в рамках Educational Data Mining, указывают на два наиболее устойчивых предиктора академической задолженности: текущую успеваемость и посещаемость [3, 4]. Эти факторы обладают двойным преимуществом: они объективно регистрируются в электронных журналах и поддаются количественному измерению без привлечения дополнительных психодиагностических или социологических инструментов, что критически важно для обеспечения массовости и оперативности прогнозирования.

Текущая успеваемость, выражаемая через средний балл за отчётный период, оценки за контрольные и домашние работы, отражает актуальный уровень освоения материала. Низкая успеваемость свидетельствует о накоплении пробелов в знаниях, которые при отсутствии корректирующих мер приводят к невозможности успешного прохождения итоговой аттестации. Особую предсказательную ценность имеет динамика оценок. Стабильно низкий или снижающийся средний балл указывает на усиливающиеся трудности, тогда как разовые неудачи могут быть случайными [3].

Посещаемость выступает вторым ключевым фактором. Систематические пропуски занятий ведут к фрагментарному усвоению учебного материала, разрыву логических связей между темами и, следовательно, к снижению успеваемости. При этом важно различать пропуски по уважительной причине и пропуски без уважительной причины. Исследования показывают, что именно доля неуважительных пропусков сопоставима с академической неуспешностью наиболее выражено [5]. Однако в условиях ограниченности данных электронных журналов, где причина пропуска не всегда оперативно фиксируется, общее количество пропусков также служит приемлемым аппроксимирующим показателем.

Помимо двух основных факторов, в литературе выделяют дополнительные: количество прошлых академических неудач, мотивация и вовлечённость учащегося, социально-экономический статус семьи, уровень образования родителей. Однако эти показатели либо труднодоступны для автоматического сбора, либо требуют специальных обследований, что ограничивает их применимость в рамках оперативного внутришкольного прогнозирования [4].

Таким образом, задача сводится к тому, чтобы на основе минимального, но достаточного набора объективных признаков – текущей успеваемости и посещаемости – выявить учащихся, чьё текущее состояние с высокой вероятностью приведёт к формальной академической задолженности. При этом важно не просто зафиксировать наличие факторов риска, но и количественно оценить их совместный вклад, что возможно с применением методов машинного обучения, рассматриваемых в следующем подразделе.

1.2 Методы машинного обучения в прогнозировании успеваемости

Задача прогнозирования академической задолженности естественным образом формулируется как бинарная классификация: по набору признаков требуется отнести учащегося к одному из двух состояний – «задолженность возникнет» или «задолженность не возникнет». Образовательные данные, как правило, представляют собой таблицы смешанного типа, в которых числовые показатели (оценки, пропуски) сочетаются с категориальными переменными. В таких условиях хорошо зарекомендовали себя как линейные методы, так и ансамблевые алгоритмы, что подтверждается практикой Educational Data Mining [3].

Логистическая регрессия даёт на выходе вероятность принадлежности к положительному классу и при этом сохраняет прозрачность. Веса при признаках прямо указывают направление и силу их влияния. Это особенно ценно в образовательном контексте, где педагогу или администратору важно понимать, почему конкретный учащийся попал в группу риска [6]. Метод ранее успешно применялся для предсказания отчислений на основе посещаемости и текущих оценок, показывая точность, сопоставимую с более сложными моделями [4].

Деревья решений и их ансамбли, в частности случайный лес, способны улавливать нелинейные комбинации признаков без ручного конструирования взаимодействий. Случайный лес строит множество деревьев на случайных подвыборках, усредняя их прогнозы, что резко снижает риск переобучения по сравнению с одиночным деревом. В задачах образовательного прогнозирования он неоднократно демонстрировал высокую обобщающую способность, причём наиболее информативными признаками стабильно оказывались именно посещаемость и текущие баллы [7, 8]. Важным побочным продуктом является ранжирование важности признаков – это даёт возможность сопоставить выводы модели с предметными ожиданиями.

Методы градиентного бустинга (XGBoost, CatBoost) потенциально могут дать небольшой выигрыш в точности за счёт последовательного исправления ошибок, однако они более чувствительны к настройке гиперпараметров и склонны к переобучению на выборках малого объёма. Исследования показывают, что при числе наблюдений порядка нескольких сотен преимущества бустинга перед случайным лесом или даже логистической регрессией часто оказываются незначительными [9].

Для настоящего исследования, в котором объём данных относительно невелик, а практическая полезность неразрывно связана с возможностью объяснить прогноз педагогу, в качестве основной модели выбрана логистическая регрессия. Случайный лес используется параллельно для контроля устойчивости результатов и проверки того, дают ли нелинейные эффекты существенное улучшение качества классификации.

1.3 Существующие системы раннего предупреждения

Задача прогнозирования академической задолженности не является новой, и в образовательной практике накоплен определённый опыт создания программных средств, направленных на снижение рисков неуспеваемости. Ниже рассмотрены два примера систем, имеющих сходные цели, но реализованных в различных контекстах, и проведено их сравнение по совокупности критериев, важных для применения в школьной среде.

Программа прогнозирования успеваемости студентов первого курса вуза по баллам ЕГЭ и текущим аттестациям. Далее – Аналог 1. Программа предназначена для поиска закономерностей между баллами вступительных испытаний, баллами текущих аттестаций и оценками за промежуточные аттестации первого курса средствами глубокого машинного обучения [10]. На выходе формируется прогноз успеваемости студента по каждой дисциплине: «зачтено» / «не зачтено» либо «отлично» / «хорошо» / «удовлетворительно» / «не удовлетворительно». Сильными сторонами данной разработки являются учёт довузовских факторов (результатов ЕГЭ) и детализированный прогноз по отдельным предметам.

Чат-бот «Цифровой помощник студента». Далее – Аналог 2. Реализованная в виде Telegram-бота, программа отслеживает дедлайны по заданиям, сопровождая напоминания мотивирующими сообщениями, и параллельно проводит диагностическое анкетирование для выявления факторов, негативно влияющих на успеваемость. На основе полученных данных строятся персонализированные программы поддержки и модели прогнозирования риска академической неуспешности [11]. Преимуществом решения является высокая вовлечённость учащихся и наличие обратной связи.

Для объективного сопоставления аналогов были выделены шесть критериев, отражающих ключевые требования к системам раннего предупреждения: эффективность (скорость обработки данных), объективность (непредвзятость), конфиденциальность (изолированность прогнозов для каждого учащегося), пользовательский опыт (удобство взаимодействия), точность (достоверность прогноза академической задолженности) и обратная связь (возможность получения прогнозов как педагогом, так и учащимся). Ниже в таблице 1 приведены экспертные оценки по десятибалльной шкале и визуальное сравнение на рисунке 1.

Таблица 1 – Экспертные оценки аналогов по выделенным критериям

Критерий	Аналог 1	Аналог 2
Эффективность	8	9
Объективность	9	9
Конфиденциальность	10	10
Пользовательский опыт	6	7
Точность	8	7
Обратная связь	7	8

Применение метода главного критерия (в качестве главного критерия выбрана «точность») показывает преимущество Аналога 1 (8 баллов против 7). Однако такая оценка является односторонней и не учитывает разную важность остальных критериев.

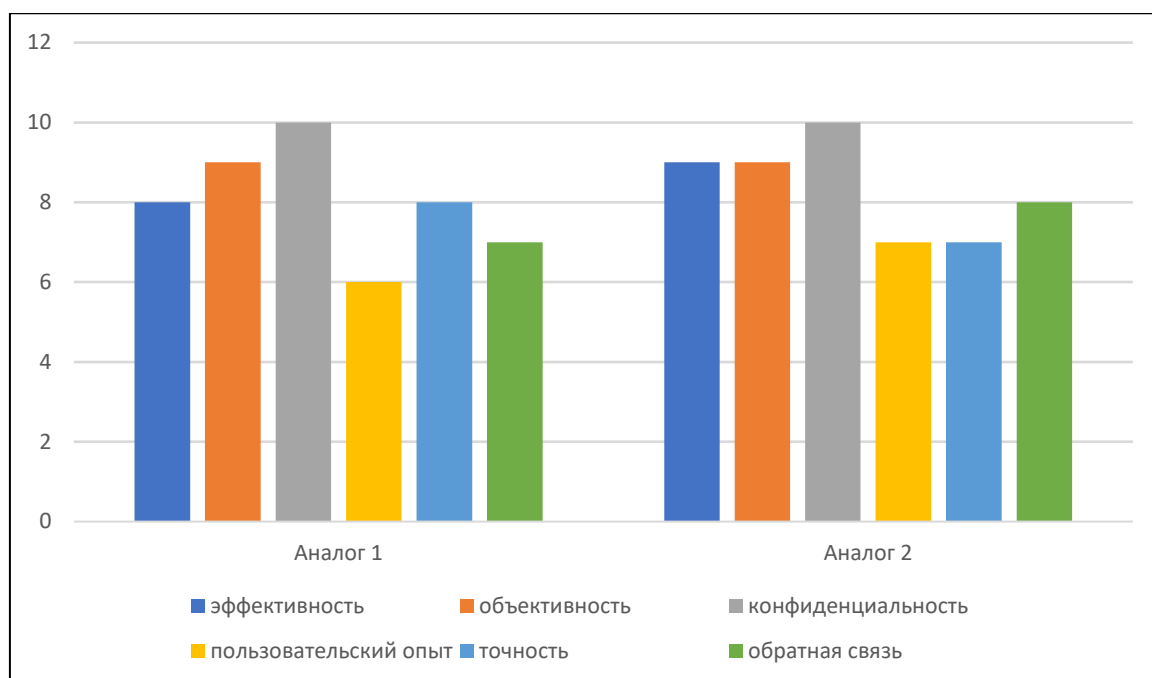


Рисунок 1 – Сравнительная диаграмма аналогов

Для более сбалансированного сравнения был использован метод аддитивной свёртки. Предварительно выполнена нормализация: по каждому критерию определено максимальное значение среди двух аналогов, затем все оценки пересчитаны как доля от максимума. Веса критериев назначены исходя из приоритетов школьной среды: наибольший вес присвоен точности (0.5), далее – пользовательскому опыту (0.12), объективности (0.1), конфиденциальности (0.06), эффективности (0.2) и обратной связи (0.02).

Нормализация – последовательность процедур, с помощью все критерии приводятся к единому масштабу измерения. В процессе нормализации выполнены следующие действия:

- Определен $\max$ каждого локального метода: $a_{j+} = \max a_{ij}$,
- В соответствии с критерием $\max$ эффективности пересчитано значение частных критериев (таблица 2).

Таблица 2 – Оценки аналогов с max

Критерий	Аналог 1	Аналог 2	Max
Эффективность	8	9	9
Объективность	9	9	9
Конфиденциальность	10	10	10
Пользовательский опыт	6	7	7
Точность	8	7	8
Обратная связь	7	8	8

Для решения данной задачи использован аддитивный критерий оптимальности (обобщенная функция цели), который определяется по формуле:

$$F(a_{ij}) = \sum_{j=1}^n \lambda_j \times a_{ij}, \quad \sum_{j=1}^m \lambda_j = 1. \quad (1)$$

где a_{ij} – значение частного (локального) критерия;

λ_j – вес (важность) j -го частного критерия.

Значения аддитивного критерия оптимальности, рассчитанные по формуле, составили $F1 = 0,684$, $F2 = 0,443$, $F3 = 0,716$.

Метод идеальной точки позволяет оценить расстояние каждого варианта до эталона, сформированного наилучшими значениями по каждому критерию (таблица 3). Визуальное сравнение расстояний представлено на рисунке 2.

$$F_i = \sqrt{(\sum_{j=1}^n (x_{ij} - x_j)^2)}. \quad (2)$$

Евклидовы расстояния, рассчитанные по формуле, составили $F_1 = 1,73$, $F_2 = 1$.

Таблица 3 – Оценки аналогов с идеальной точкой

Критерий	Аналог 1	Аналог 2	Идеальная точка
Эффективность	8	9	9
Объективность	9	9	9
Конфиденциальность	10	10	10
Пользовательский опыт	6	7	7
Точность	8	7	8
Обратная связь	7	8	8

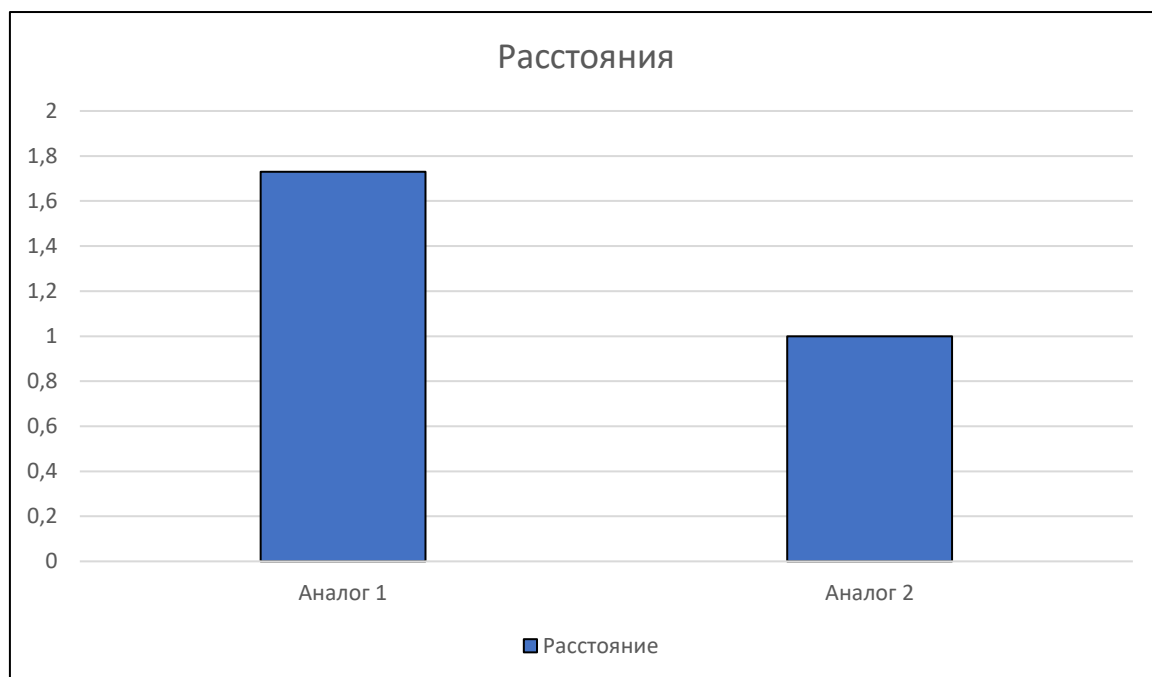


Рисунок 2 – График расстояний

1.4 Выводы по обзору литературы

Анализ литературы и существующих систем, проведённый в настоящей главе, позволяет сформулировать несколько взаимосвязанных заключений, которые в совокупности обосновывают направление и методологию дальнейшего исследования.

Прежде всего, следует подчеркнуть, что академическая задолженность как образовательный феномен имеет достаточно чётко очерченные и эмпирически подтверждённые предикторы. Нормативные документы фиксируют прямую связь между неудовлетворительными результатами промежуточной аттестации, систематическими пропусками занятий и возникновением задолженности [1]. Эта правовая конструкция подкрепляется

данными многочисленных эмпирических работ, выполненных в русле Educational Data Mining. В них показано, что текущая успеваемость и посещаемость являются наиболее устойчивыми и универсальными факторами, позволяющими с приемлемой точностью прогнозировать дальнейшие академические трудности, независимо от того, снижение ли это итоговых оценок, возникновение задолженности или отчисление [3, 5]. Принципиально важно, что оба показателя носят объективный характер, автоматически фиксируются в электронных журналах и не требуют проведения дополнительных обследований, опросов или экспертных оценок. Это делает их идеальными кандидатами на роль базовых предикторов в системах раннего оповещения, ориентированных на массовое применение в школьной среде.

Обращение к методам машинного обучения, применяемым в задачах образовательного прогнозирования, обнаруживает несколько отчётливых тенденций. Логистическая регрессия на протяжении многих лет остаётся одним из наиболее востребованных инструментов именно в силу своей прозрачности: весовые коэффициенты при признаках поддаются прямой содержательной интерпретации, что особенно ценно в педагогическом контексте, где важно не просто получить прогноз, но и объяснить его учителю или родителю [3, 4]. Случайный лес, со своей стороны, зарекомендовал себя как надёжный метод, способный улавливать нелинейные зависимости и взаимодействия между признаками без ручного конструирования этих взаимодействий [8]. При этом сопоставление двух подходов на образовательных данных нередко показывает, что выигрыш в точности от применения ансамблевых моделей не является кардинальным, тогда как потери в интерпретируемости могут быть существенными. Более сложные алгоритмы, такие как градиентный бустинг, демонстрируют преимущества в первую очередь на больших выборках с разнородными признаками; при малом объёме данных их отрыв от линейных моделей минимален, а риск переобучения возрастает [9]. Поэтому для условий, в которых количество

наблюдений не превышает нескольких сотен, а практическая ценность неразрывно связана с возможностью дать педагогу понятное обоснование прогноза, логистическая регрессия представляется оправданным выбором в качестве основной модели, тогда как случайный лес может быть использован для проверки устойчивости результатов и оценки потенциального вклада нелинейных эффектов.

Рассмотрение существующих программных средств, нацеленных на снижение академической неуспешности, показало, что на сегодняшний день они ориентированы главным образом на высшее образование. Проанализированные системы – программа прогнозирования успеваемости студентов первого курса по баллам ЕГЭ и чат-бот «Цифровой помощник студента» – опираются на специфические для вуза показатели (баллы вступительных испытаний, дедлайны по заданиям, данные анкетирования) и не используют в явном виде такой важнейший для школы маркер, как регулярность посещения занятий [10, 11]. Многокритериальное сравнение аналогов, выполненное с применением метода главного критерия, аддитивной свёртки и метода идеальной точки, выявило, что в зависимости от расстановки приоритетов преимущество может получать то одна, то другая система, однако ни одна из них не сочетает одновременно высокую точность прогноза, простоту внедрения и опору на минимальный, повсеместно доступный набор данных – текущие оценки и пропуски.

Совокупность изложенных обстоятельств формирует отчётливую исследовательскую нишу: существует потребность в модели, которая была бы сконструирована специально под условия общеобразовательной школы, работала бы с теми данными, которые уже имеются в электронных журналах, и при этом сохраняла бы высокую интерпретируемость результатов. Именно на заполнение этой ниши направлено настоящее исследование. В следующей главе описываются материалы и методы, использованные для построения и оценки прогностического прототипа, в основу которого положен открытый датасет Student Performance.

2 Материалы и методы прогнозирования

2.1 Описание датасета

Исходным материалом для настоящего исследования послужил открытый набор данных Student Performance, созданный Паулу Кортесом и Алисой Силвой в 2008 году и размещённый в репозитории машинного обучения Университета Калифорнии в Ирвайне (UCI Machine Learning Repository) [12]. Датасет содержит информацию об успеваемости учащихся двух португальских средних школ – имени Габриэла Перейры и имени Моузинью да Силвейры – по двум основным дисциплинам: математике и португальскому языку. Данные собирались путём выгрузки из школьных информационных систем и дополнялись анкетными сведениями, что обеспечило сочетание объективных академических показателей с демографическими и социальными характеристиками учащихся.

Полный датасет разделён на два файла: student-mat.csv (математика) и student-por.csv (португальский язык), каждый из которых содержит 395 и 649 записей соответственно. Настоящее исследование построено на файле student-mat.csv, включающем 395 записей и 33 переменные. Выбор именно этого файла обусловлен тем, что математика традиционно относится к дисциплинам с наиболее высоким уровнем академических трудностей у школьников, а значит, задача прогнозирования задолженности по этому предмету является особенно актуальной.

Каждая запись датасета соответствует одному учащемуся и содержит полный набор характеристик, зафиксированных на момент окончания учебного года. Переменные могут быть сгруппированы в несколько содержательных категорий: демографические данные, социально-семейные факторы, школьные характеристики, текущая успеваемость и посещаемость. В таблице 4 приведены описания переменных, непосредственно задействованных в настоящем исследовании или важных для понимания структуры данных.

Таблица 4 – Основные переменные датасета student-mat.csv

Переменная	Тип данных	Описание	Пример значения
School	Категориальная	Школа (GP – Габриэл Перейра, MS – Моузинью да Силвейра)	GP
Sex	Категориальная	Пол учащегося	F
Age	Числовая	Возраст, лет	16
Studytime	Числовая	Еженедельное время на учёбу (1 – <2 ч., 2 – 2-5 ч., 3 – 5-10 ч., 4 – >10 ч.)	2
Failures	Числовая	Количество несданных экзаменов в прошлые годы (0–3)	0
Absences	Числовая	Количество пропущенных уроков за год	6
G1	Числовая	Оценка за первый семестр (0–20)	11
G2	Числовая	Оценка за второй семестр (0–20)	10
G3	Числовая	Итоговая оценка за год (0–20)	11

Помимо перечисленных, датасет содержит и другие переменные, которые в рамках данной работы не использовались, но создают широкий контекст для понимания природы данных: адрес проживания (городской/сельский), размер семьи, уровень образования и профессия родителей, наличие дополнительной образовательной поддержки, потребление алкоголя в учебные и выходные дни и другие.

Система оценивания, принятая в португальских школах, использует двадцатибалльную шкалу, где минимальным порогом, отделяющим удовлетворительную оценку от неудовлетворительной, традиционно считается 10 баллов. Это послужило основанием для выбора именно этой границы при формировании целевой переменной. Оценки G1 и G2 отражают успеваемость в первом и втором семестрах соответственно и могут рассматриваться как прокси текущей успеваемости на разных этапах учебного

года, тогда как оценка G3 представляет собой итоговый результат, на основе которого и определяется наличие или отсутствие задолженности.

Распределение учащихся по школам в датасете неравномерно: примерно 88% записей относятся к школе имени Габриэла Перейры и 12% – к школе имени Моузинью да Силвейры. Это обстоятельство не создаёт принципиальных препятствий для анализа, однако должно учитываться при интерпретации результатов, поскольку выборка не является строго репрезентативной относительно генеральной совокупности.

Среднее значение итоговой оценки G3 составляет около 10.4 балла при стандартном отклонении порядка 4.6 балла, что говорит о заметном разбросе успеваемости и наличии как весьма успешных, так и испытывающих серьёзные трудности учащихся. Количество пропусков варьируется от 0 до 75, при этом медианное значение составляет 4 пропуска за год, что указывает на сильную правостороннюю асимметрию распределения: основная масса учеников посещает занятия достаточно регулярно, однако присутствует небольшое число учащихся с экстремально высокой неявкой.

Выбор данного датасета для проверки гипотезы исследования обоснован несколькими соображениями. Он содержит как оценки за различные периоды, так и сведения о пропусках, что в точности соответствует выделенным в главе 1 ключевым факторам риска. Его структура близка к тому набору данных, который может быть извлечён из типового школьного электронного журнала. Такие как идентификатор учащегося, оценки по предмету и информация о посещаемости. Открытый характер датасета обеспечивает воспроизводимость результатов. Любой исследователь может повторить описанные ниже вычислительные эксперименты и получить те же выводы. Также можно отметить достаточный объём выборки (395 записей), позволяющий сосредоточиться на содержательной стороне моделирования, не отвлекаясь на технические аспекты работы с большими данными.

2.2 Подготовка данных и конструирование признаков

Исходный датасет `student-mat.csv` поступил в анализ в предварительно очищенном виде. Пропущенные значения в нём отсутствуют, типы переменных согласованы. Поэтому процедуры заполнения пропусков и нормализации на этапе подготовки не потребовались. Основное внимание было уделено созданию производных признаков, способных компактно отразить ключевые факторы, выделенные в обзоре литературы, такие как текущая успеваемость и уровень посещаемости, а также формированию целевой переменной, соответствующей формальному определению академической задолженности.

В исходных данных успеваемость представлена двумя переменными: G1 (оценка за первый семестр) и G2 (оценка за второй семестр). Поскольку оба показателя отражают актуальный уровень подготовленности учащегося перед итоговой аттестацией, было принято решение объединить их в единый показатель `avg_grade` – средний балл за период:

$$avg_grade = (G1 + G2) / 2. \quad (3)$$

где G1 – оценка за первый семестр;

G2 – оценка за второй семестр.

Такое агрегирование сглаживает случайные колебания успеваемости и даёт более устойчивую оценку, чем каждый из семестровых показателей по отдельности. В отличие от итоговой оценки G3, которая формируется по завершении года и используется только как целевая переменная, средний балл за два семестра доступен до итоговой аттестации и может служить предиктором.

Далее рассмотрим посещаемость и конструирование риск-скора. Переменная `absences` содержит общее количество пропущенных уроков за год. Предварительный корреляционный анализ показал, что `absences` отрицательно

коррелирует с итоговой оценкой G3. Переменная *failures*, фиксирующая число прошлых академических неудач, продемонстрировала ещё более выраженную отрицательную связь, что соответствует ожиданиям. Напротив, текущая успеваемость *avg_grade* положительно связана с итоговым результатом. Для учёта совместного влияния этих трёх факторов был сконструирован композитный показатель – риск-скор (*risk_score*):

$$risk_score = absences \times 0.4 + failures \times 3 - avg_grade \times 1.5. \quad (4)$$

где *absences* – количество пропусков;

failures – число прошлых академических неудач;

avg_grade – текущая успеваемость.

Весовые коэффициенты подобраны эмпирически, исходя из нормированных коэффициентов корреляции трёх переменных с G3, и отражают следующий принцип: пропуски и прошлые неудачи увеличивают риск задолженности (знак плюс), тогда как высокая текущая успеваемость снижает его (знак минус). Абсолютные величины коэффициентов призваны уравновесить масштабы переменных и придать наибольший вес фактору прошлых неудач, который как в литературе, так и в данных демонстрирует наиболее сильную связь с итоговым результатом обучения.

Переменная *studytime* (еженедельное время, уделяемое учёбе, в дискретной шкале от 1 до 4) была оставлена в исходном виде как дополнительный предиктор. Несмотря на то, что этот показатель основан на самооценке учащегося и может содержать элемент субъективности, он даёт полезную информацию о вовлечённости в учебный процесс и, как показали предшествующие исследования [4; 8], нередко вносит положительный вклад в прогноз.

Поэтому базовый набор признаков для моделирования был сформирован из двух переменных: *risk_score* и *studytime*. Риск-скор интегрирует

информацию о посещаемости, прошлых неудачах и текущей успеваемости в единую метрику, а studytime дополняет её поведенческим аспектом.

Чтобы оценить, насколько полезным оказалось конструирование синтетического риск-скора, параллельно был подготовлен альтернативный набор признаков, включающий исходные переменные без предварительной агрегации: absences, failures, avg_grade и studytime. Сравнение моделей, обученных на базовом и альтернативном наборах, позволяет судить о том, теряется ли предсказательная сила при переходе от синтетического показателя к разреженному набору отдельных факторов.

Формальное определение академической задолженности предполагает неудовлетворительный результат итоговой аттестации. В португальской двадцатибалльной системе порогом, отделяющим удовлетворительную оценку от неудовлетворительной, традиционно считается 10 баллов. В соответствии с этим итоговая оценка G3 была бинаризована: debt = 1, если G3 <= 10, и debt = 0 в противном случае.

Исходя из этого, класс 1 соответствует наличию задолженности по предмету, класс 0 – её отсутствию. Порог в 10 баллов является минимальным допустимым значением для получения зачёта по дисциплине в большинстве образовательных систем, использующих 20-балльную шкалу.

Поскольку в базовом наборе оба признака имеют один порядок величин, а в альтернативном наборе переменные absences, failures, avg_grade и studytime могут различаться по масштабу, для чистоты последующего сравнения было проверено влияние стандартизации. Однако, как показали предварительные прогоны, она не оказала значимого влияния на коэффициенты логистической регрессии (в силу встроенной регуляризации) и на качество классификации. В связи с этим обучение моделей проводилось на немасштабированных данных, что упрощает интерпретацию коэффициентов и воспроизведение результатов.

Сформированные описанным образом признаки и целевая переменная послужили входом для процедуры моделирования.

2.3 Методы моделирования

Для решения поставленной задачи бинарной классификации были выбраны два алгоритма, представляющих различные классы методов машинного обучения: логистическая регрессия как базовая линейная модель и случайный лес как ансамблевый метод, способный улавливать нелинейные зависимости. Выбор именно этих алгоритмов обоснован в обзоре литературы и обусловлен, с одной стороны, требованием интерпретируемости результатов, с другой – необходимостью проверить, дают ли нелинейные эффекты значимое улучшение качества прогноза на имеющемся датасете.

Логистическая регрессия моделирует вероятность принадлежности наблюдения к классу 1 (наличие задолженности) как сигмоидную функцию от линейной комбинации признаков. В данном исследовании использовалась реализация `LogisticRegression` из библиотеки `scikit-learn` с параметрами по умолчанию, за исключением увеличенного до 1000 числа итераций (`max_iter = 1000`), что гарантировало сходимость оптимизационного алгоритма на обучающей выборке. Регуляризация применялась в форме L2 (гребневая), что помогает избежать переобучения на малых данных и стабилизирует оценки коэффициентов. Линейный характер модели означает, что вклад каждого признака в итоговый прогноз аддитивен и не зависит от значений других признаков. Это ограничение компенсируется тем, что один из признаков – риск-скор – уже является агрегированным показателем, объединяющим несколько исходных переменных с весами, отражающими их совместное влияние. Поэтому линейная модель, обученная на риск-скор, опосредованно учитывает и нелинейные эффекты, заложенные в конструкцию признака. К преимуществам логистической регрессии в контексте данной задачи относится не только прозрачность, но и возможность прямой интерпретации знака и абсолютной величины коэффициентов как показателей направления и силы влияния каждого фактора.

Случайный лес представляет собой ансамбль деревьев решений, каждое из которых строится по случайной подвыборке наблюдений и случайному подмножеству признаков. Итоговое предсказание формируется путём

усреднения вероятностей, выданных отдельными деревьями (мягкое голосование). В работе применялся классификатор `RandomForestClassifier` библиотеки `scikit-learn` также со стандартными гиперпараметрами: количество деревьев – 100, максимальная глубина не ограничена, минимальное число наблюдений в листе – 1 для классификации. Такой выбор оправдан в условиях малой выборки, где ансамбль из сотни полных деревьев способен детально описать структуру данных без существенного риска переобучения, поскольку случайный характер подвыборок и подмножеств признаков обеспечивает декорреляцию отдельных деревьев и, как следствие, низкую дисперсию итогового прогноза. Случайный лес использовался в исследовании как сравнительный ориентир. Если его метрики качества окажутся значительно выше, чем у логистической регрессии, это укажет на присутствие сильных нелинейных взаимодействий, которые линейная модель не улавливает. Кроме того, случайный лес предоставляет встроенный механизм оценки важности признаков на основе среднего снижения примеси, что позволяет сопоставить его выводы с коэффициентами логистической регрессии.

Обе модели были реализованы на языке Python 3 с использованием библиотек `pandas` (версия 1.5.3), `numpy` (1.24.3) и `scikit-learn` (1.2.2). Вычисления проводились в облачной среде Google Colab, что обеспечило воспроизводимость эксперимента при отсутствии требований к локальному аппаратному обеспечению. Полный код прототипа, включая загрузку данных, предобработку, обучение и визуализацию результатов, приведён в Приложении.

2.4 Процедура обучения и оценки

Обучение и оценка качества моделей проводились по стандартной схеме, применяемой в задачах бинарной классификации на табличных данных достаточного объёма.

Исходный датасет, содержащий 395 наблюдений, был разбит на обучающую и тестовую выборки в соотношении 80% / 20%. Разбиение выполнялось функцией `train_test_split` библиотеки `scikit-learn` с

фиксированным начальным состоянием генератора псевдослучайных чисел (`random_state = 42`), что гарантирует точную воспроизводимость экспериментов. В силу относительно сбалансированного распределения целевой переменной дополнительная стратификация не применялась, однако при необходимости она может быть легко добавлена в последующих итерациях. Обучающая выборка, таким образом, составила 316 записей, тестовая – 79 записей. Этот объём достаточен для построения устойчивой линейной модели и ансамбля деревьев, но при этом оставляет достаточно примеров для объективной проверки предсказательной способности на ранее не виданных данных.

Поскольку задача формулируется как бинарная классификация с практическим приоритетом обнаружения учащихся группы риска (класс 1), для оценки качества использовался набор метрик, каждая из которых отражает определённый аспект поведения модели.

Матрица ошибок (*confusion matrix*) даёт наиболее полное представление о результатах классификации, разделяя все предсказания на четыре категории: истинно положительные (TP – верно предсказанная задолженность), истинно отрицательные (TN – верно предсказанное отсутствие задолженности), ложно положительные (FP – ложная тревога, когда благополучный учащийся ошибочно помечен как рискованный) и ложно отрицательные (FN – пропуск реальной задолженности, наиболее нежелательная ошибка). Матрица ошибок визуализировалась и анализировалась для каждой модели.

На основе матрицы ошибок рассчитывались производные метрики:

– Accuracy – доля правильных ответов в эксперименте. Общая точность модели, вычисляется по формуле:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN). \quad (5)$$

где TP – верно предсказанная задолженность;

TN – верно предсказанное отсутствие задолженности;

FP – ложная тревога;

FN – пропуск реальной задолженности.

Эта метрика удобна как обобщённый индикатор, но при несбалансированных классах может вводить в заблуждение, поэтому она дополнялась другими показателями,

– Precision – точность для класса 1. Доля истинно определенных задолженников среди всех, кого модель отнесла к этому классу: $TP / (TP + FP)$. Высокое значение precision означает, что модель редко ошибается, предсказывая задолженность, но ничего не говорит о том, какую долю реальных задолженностей она пропускает,

– Recall – полнота для класса 1. Доля истинно определенных задолженников, которых модель обнаружила: $TP / (TP + FN)$. Эта метрика в контексте исследования является приоритетной, поскольку пропуск учащегося с реальным риском задолженности влечёт более серьёзные последствия, чем ложная тревога. Максимизация recall означает, что модель стремится найти как можно больше потенциальных должников,

– F1-мера – гармоническое среднее показателей precision и recall: $2 \cdot (Precision \cdot Recall) / (Precision + Recall)$. Эта метрика даёт сбалансированную оценку и используется в условиях, когда одновременно важны и точность, и полнота, но классы не слишком разбалансированы.

Перечисленные метрики вычислялись средствами библиотеки scikit-learn: `accuracy_score`, `confusion_matrix`, `classification_report`. Отчёт о классификации формировался для каждой модели и включал precision, recall и F1 для обоих классов, а также сводные показатели.

Для логистической регрессии мерой важности признаков служат весовые коэффициенты модели (`model.coef_`). Положительный знак коэффициента означает, что рост соответствующего признака увеличивает вероятность отнесения к классу «задолженность», отрицательный – снижает её. Абсолютная величина коэффициента указывает на силу влияния при прочих равных условиях. Для случайного леса использовался встроенный

атрибут `feature_importances_`, вычисляемый на основе среднего снижения примеси. Сопоставление двух наборов важностей позволяет проверить, согласуются ли выводы разных моделей относительно ключевых факторов риска. Никакой специальной процедуры подбора гиперпараметров не проводилось, поскольку цель прототипа состояла не в максимизации метрик, а в проверке принципиальной предсказательной способности выбранных факторов. Модели сравнивались по значениям `recall`, `precision` и F1-меры, полученным на тестовой выборке, а также по содержательной интерпретируемости результатов.

Описанная процедура обучения и оценки обеспечивает прозрачность эксперимента и его полную воспроизводимость с использованием только открытых данных и общедоступных программных средств.

2.5 Инструментарий

Все вычислительные эксперименты, включая загрузку данных, предобработку, обучение моделей и визуализацию результатов, были выполнены на языке Python 3 (версия 3.10.12) в облачной среде Google Colab. Выбор этой платформы обусловлен её общедоступностью, отсутствием необходимости в установке программного обеспечения на локальный компьютер и возможностью точного воспроизведения результатов другими исследователями.

Для работы с данными использовались библиотеки:

- `pandas` (версия 1.5.3) – чтение, обработка и агрегация табличных данных,
- `numpy` (версия 1.24.3) – выполнение векторных и матричных операций,
- `matplotlib` (версия 3.7.1) – построение всех графиков и визуализаций, представленных в главе 3.

Модели машинного обучения были реализованы средствами библиотеки `scikit-learn` (версия 1.2.2), которая предоставляет унифицированный интерфейс

для разделения выборок, обучения классификаторов и расчёта метрик качества. В частности, применялись:

- `train_test_split` – разбиение данных на обучающую и тестовую выборки,
- `LogisticRegression` – построение модели логистической регрессии,
- `RandomForestClassifier` – построение модели случайного леса,
- `accuracy_score`, `confusion_matrix`, `classification_report` – вычисление метрик и формирование отчётов о классификации.

Полный исходный код прототипа, включая загрузку датасета, расчёт признаков, обучение обеих моделей и генерацию всех иллюстраций, приведён в Приложении А. Все промежуточные выводы, представленные в последующих разделах, могут быть верифицированы повторным запуском этого кода в любой совместимой среде.

2.6 Выводы по главе 2

В настоящей главе была сформирована методологическая основа исследования, охватывающая все этапы работы с данными – от их происхождения и структуры до конкретных процедур обучения и оценки прогнозных моделей.

Исходным материалом послужил открытый датасет *Student Performance*, содержащий сведения об успеваемости и посещаемости 395 учащихся португальских школ по математике. Выбор именно этого набора данных продиктован его содержательным соответствием целям исследования: в нём одновременно присутствуют и текущие оценки за два семестра, и итоговая оценка, и показатели пропусков, что в точности воспроизводит минимально необходимый набор признаков, доступный в типовом электронном журнале. Двадцатибалльная система оценивания с устоявшимся порогом удовлетворительной оценки в 10 баллов обеспечила естественное и прозрачное определение целевой переменной.

На этапе подготовки данных основное внимание было уделено конструированию признаков, в сжатой форме выражающих ключевые

факторы риска, выделенные в обзоре литературы. Средний балл за два семестра позволил агрегировать текущую успеваемость в единый показатель, а композитный риск-скор объединил сведения о пропусках, прошлых неудачах и текущей успеваемости с весами, эмпирически согласованными с выявленными корреляционными связями. Параллельно был предусмотрен альтернативный набор признаков из исходных переменных без предварительной агрегации, что заложило основу для последующего сопоставления предсказательной силы разреженного и синтетического представлений.

В качестве инструментов моделирования были выбраны два алгоритма, соответствующие критериям интерпретируемости и устойчивости на малых выборках: логистическая регрессия и случайный лес. Такой выбор не случаен – он опирается на анализ предшествующих работ в области Educational Data Mining и намеренно избегает неоправданного усложнения, позволяя сосредоточиться на содержательной проверке гипотезы, а не на гонке за десятками долями процента точности. Оценка качества моделей была организована вокруг метрик, чувствительных к задаче раннего выявления учащихся группы риска, с особым вниманием к полноте, что отражает практический приоритет – минимизацию числа незамеченных потенциальных должников.

Все вычислительные шаги реализованы на общедоступном стеке технологий (Python, scikit-learn, облачная среда Google Colab), что обеспечивает полную воспроизводимость результатов и возможность независимой верификации.

На данном этапе методологический фундамент исследования подготовлен, применённые процедуры и инструменты обоснованы, и можно перейти к изложению полученных результатов – разведочного анализа данных и итогов моделирования.

3 Анализ результатов исследования

3.1 Разведочный анализ данных

До построения прогнозных моделей был выполнен разведочный анализ датасета, преследовавший две цели: во-первых, выявить общие закономерности в распределении ключевых переменных, во-вторых, проверить наличие и характер связей между потенциальными предикторами и итоговой оценкой. Результаты анализа представлены в виде серии графиков, последовательно раскрывающих различные аспекты данных.

Первый шаг состоял в изучении распределения количества пропусков. Соответствующая гистограмма приведена на рисунке 3. Распределение обладает выраженной правосторонней асимметрией: медианное значение составляет 4 пропуска за учебный год, в то время как основная масса учащихся (более 75%) пропускает не более 7 уроков. Вместе с тем наблюдаются единичные экстремальные значения – 30, 50 и даже 75 пропусков. Наличие таких наблюдений, с одной стороны, подтверждает, что в выборке присутствуют учащиеся с очевидными признаками учебной дезадаптации, с другой – ставит вопрос о приспособленности модели к выбросам. При этом искусственно исключать эти записи нецелесообразно, поскольку именно они могут представлять наибольший интерес с точки зрения прогнозирования задолженности.

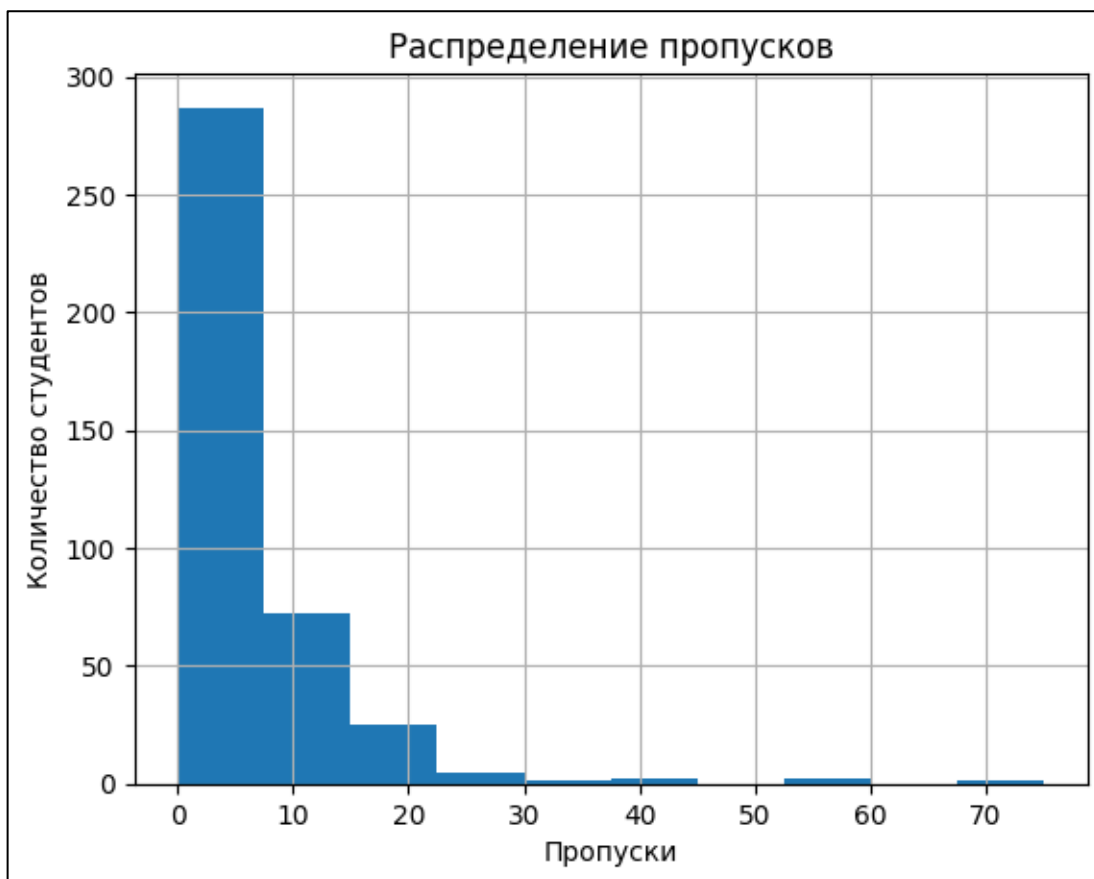


Рисунок 3 – Распределение количества пропусков (absences)

Для первичной оценки связи между пропусками и итоговой успеваемостью была построена диаграмма рассеяния, отображающая зависимость G3 от absences, отображенная на рисунке 4. Визуально прослеживается слабая отрицательная тенденция: с ростом числа пропусков итоговая оценка в среднем снижается. Однако разброс точек остаётся весьма значительным – встречаются как учащиеся с большим количеством пропусков, получившие итоговую оценку выше средней, так и ученики с единичными пропусками, но крайне низкой G3. Это указывает на то, что пропуски являются значимым, но не единственным фактором, определяющим академический результат. Данный вывод хорошо согласуется с положениями, рассмотренными в обзоре литературы, где подчёркивалась многофакторная природа академической задолженности.

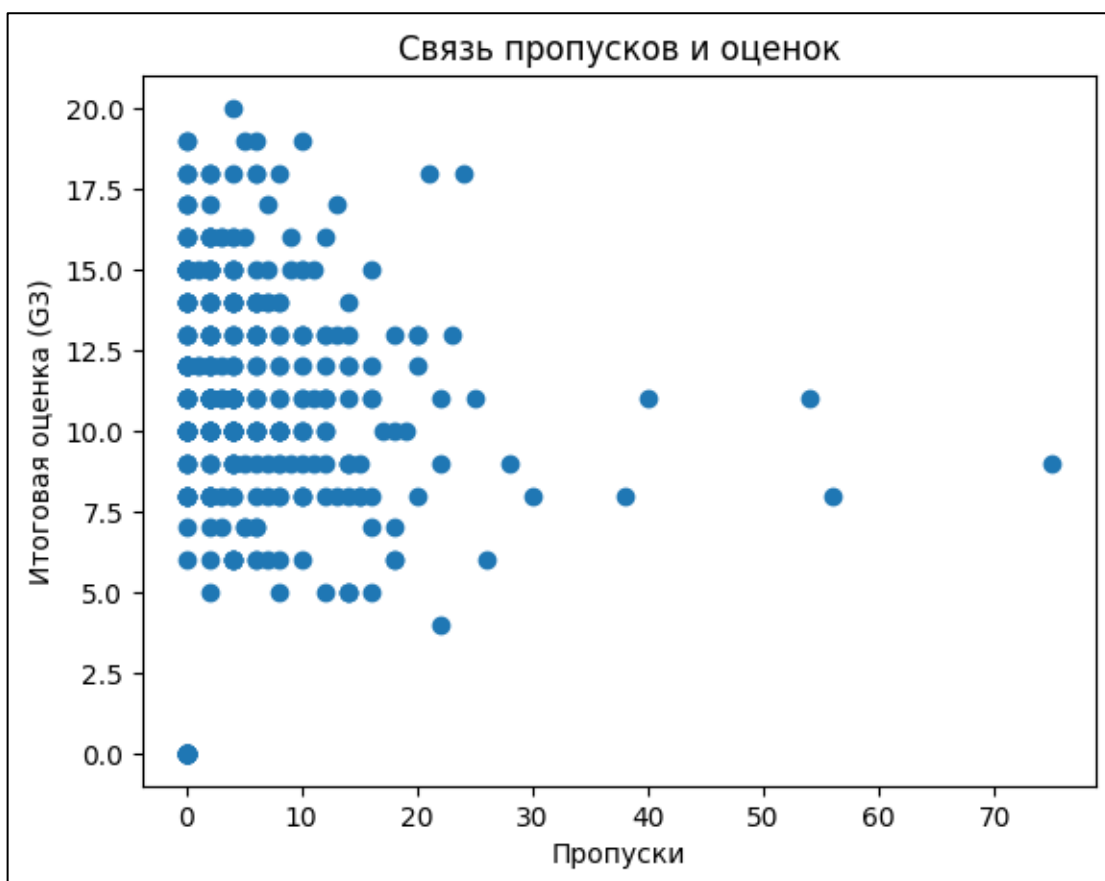


Рисунок 4 – Диаграмма рассеяния absences и итоговой оценки G3

Целевая переменная *debt*, образованная бинаризацией G3 по порогу 10 баллов, распределилась следующим образом: 256 учащихся (64,8%) не имеют задолженности ($debt = 0$), 139 учащихся (35,2%) имеют задолженность ($debt = 1$). На рисунке 5 представлена столбчатая диаграмма распределения классов. Некоторый дисбаланс в пользу успевающих учащихся присутствует, но не является критичным: модель, предсказывающая отсутствие задолженности для всех наблюдений, дала бы ассурасу порядка 65%, что задаёт реалистичный базовый уровень, который содержательная модель должна превзойти. При этом, как отмечалось ранее, приоритетной метрикой остаётся полнота выявления класса «задолженность», а не общая точность.

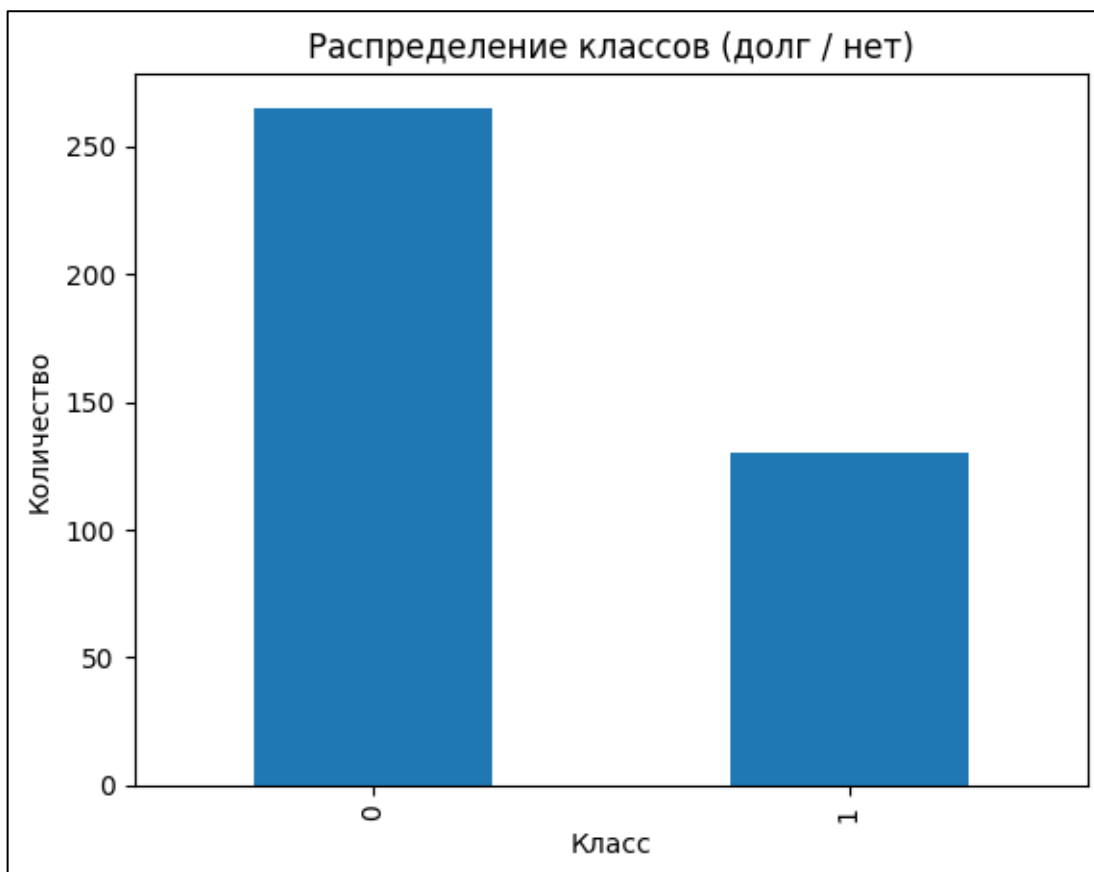


Рисунок 5 – Распределение целевой переменной debt

Чтобы получить более полное представление о взаимосвязях между числовыми переменными датасета, была построена корреляционная матрица, визуализированная в виде тепловой карты (рисунок 6). В анализ были включены оценки за первый и второй семестры (G1, G2), итоговая оценка (G3), количество пропусков (absences), число прошлых академических неудач (failures) и время, уделяемое учёбе (studytime). Корреляционная матрица обнаруживает несколько характерных закономерностей. Наиболее сильная положительная связь наблюдается между оценками разных периодов: коэффициент корреляции между G2 и G3 составляет около 0,90, между G1 и G2 – 0,85. Это свидетельствует о высокой инерционности успеваемости: учащиеся, хорошо учившиеся в первом семестре, с высокой вероятностью сохраняют тенденцию и в дальнейшем. Пропуски демонстрируют умеренную отрицательную корреляцию с G3 (порядка $-0,15 \dots -0,20$), а показатель failures – ещё более выраженную отрицательную связь ($-0,20 \dots -0,30$). Переменная studytime, напротив, показывает слабую положительную связь с итоговой

оценкой (около +0,10 ... +0,15). Ни одна пара признаков не демонстрирует экстремально высокой взаимной корреляции, которая могла бы вызвать проблему мультиколлинеарности при использовании в линейной модели. В целом структура корреляций подтверждает обоснованность отбора переменных, положенных в основу сконструированной оценки риска.

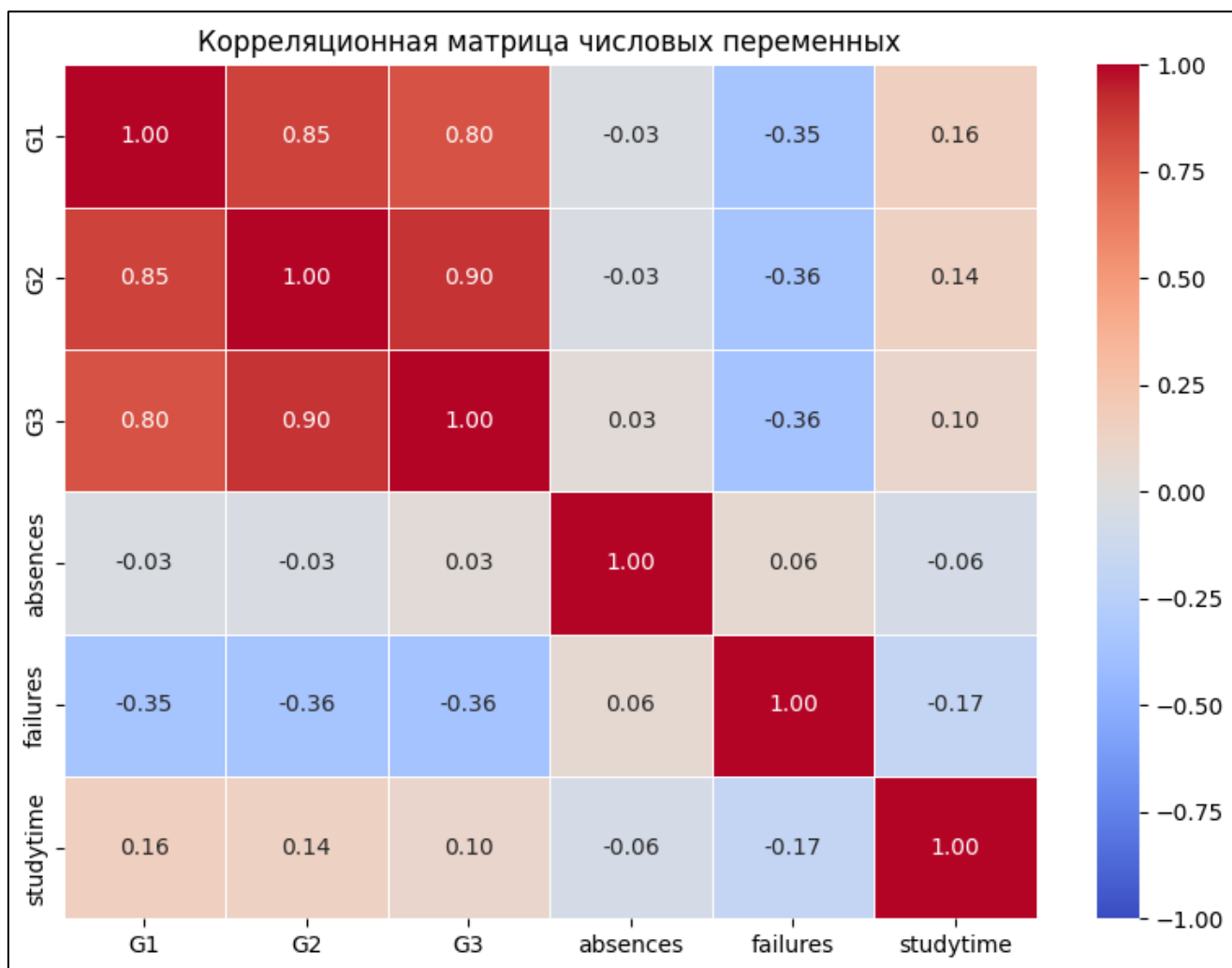


Рисунок 6 – Корреляционная матрица числовых переменных

После этого был проведён краткий анализ различий в успеваемости в зависимости от категориальных переменных, таких как школа и причины её выбора. На рисунке 7 представлены ящичковые диаграммы распределения итоговой оценки G3 для двух школ: имени Габриэла Перейры (GP) и имени Моузинью да Силвейры (MS). Медианные значения G3 в обеих школах близки (около 11 баллов), однако межквартильный размах в школе GP несколько шире, что указывает на большее разнообразие образовательных траекторий в

этой подвыборке. Это различие может быть связано с социальным составом учащихся или особенностями преподавания, однако его масштаб не настолько велик, чтобы требовать построения отдельных моделей для каждой школы. Аналогичный анализ по переменной reason (причина выбора школы: близость к дому, репутация, предпочтительная учебная программа и др.) не выявил статистически значимых различий в итоговой успеваемости между группами, хотя отдельные категории могут быть интересны для более детального изучения в будущем.

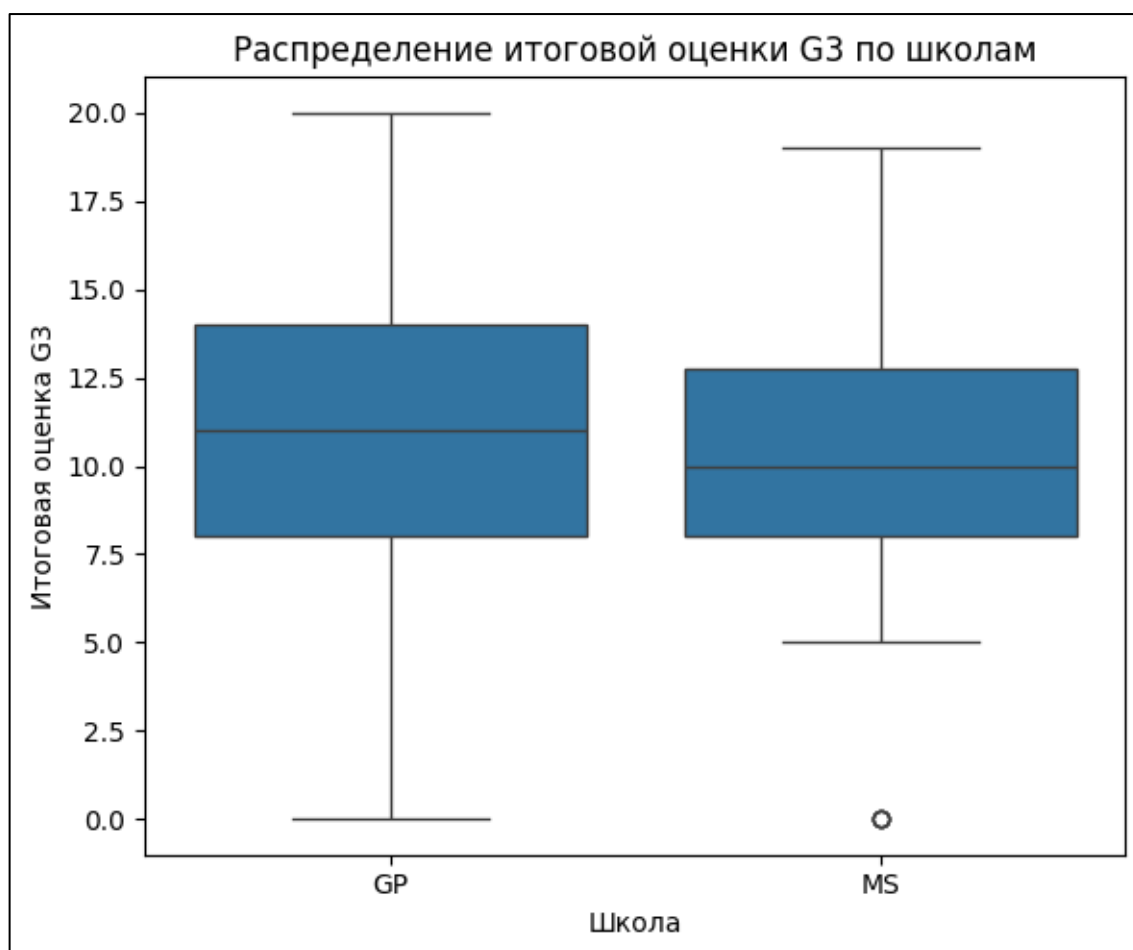


Рисунок 7 – Распределение итоговой оценки G3 по школам

Опираясь на это, можно утверждать, что разведочный анализ подтвердил наличие в данных ожидаемых закономерностей: пропуски и прошлые неудачи отрицательно связаны с итоговой оценкой, текущая успеваемость отличается высокой устойчивостью, а дисбаланс классов не

является критическим. Полученные выводы служат эмпирической основой для перехода к построению и оценке прогнозных моделей, результаты которых будут описаны далее.

3.2 Результаты моделирования логистической регрессии

После подготовки данных и конструирования признаков, была обучена модель логистической регрессии на обучающей выборке с двумя предикторами: `risk_score` и `studytime`. Целевая переменная – `debt` (бинарный признак наличия задолженности по итогам года). Обучение проводилось с параметрами, приведёнными в разделе 2.3, на 316 наблюдениях; тестирование – на 79 отложенных записях.

На рисунке 8 представлена матрица ошибок, полученная на тестовой выборке. По главной диагонали отображены наблюдения, классифицированные верно: 47 учащихся без задолженности (True Negative) и 22 учащихся с задолженностью (True Positive). Ложные тревоги (False Positive) составили 4 случая – учащиеся, ошибочно отнесённые моделью к группе риска. Пропуски реальных задолженностей (False Negative) – 6 случаев. Таким образом, модель правильно распознала большинство объектов обеих категорий.

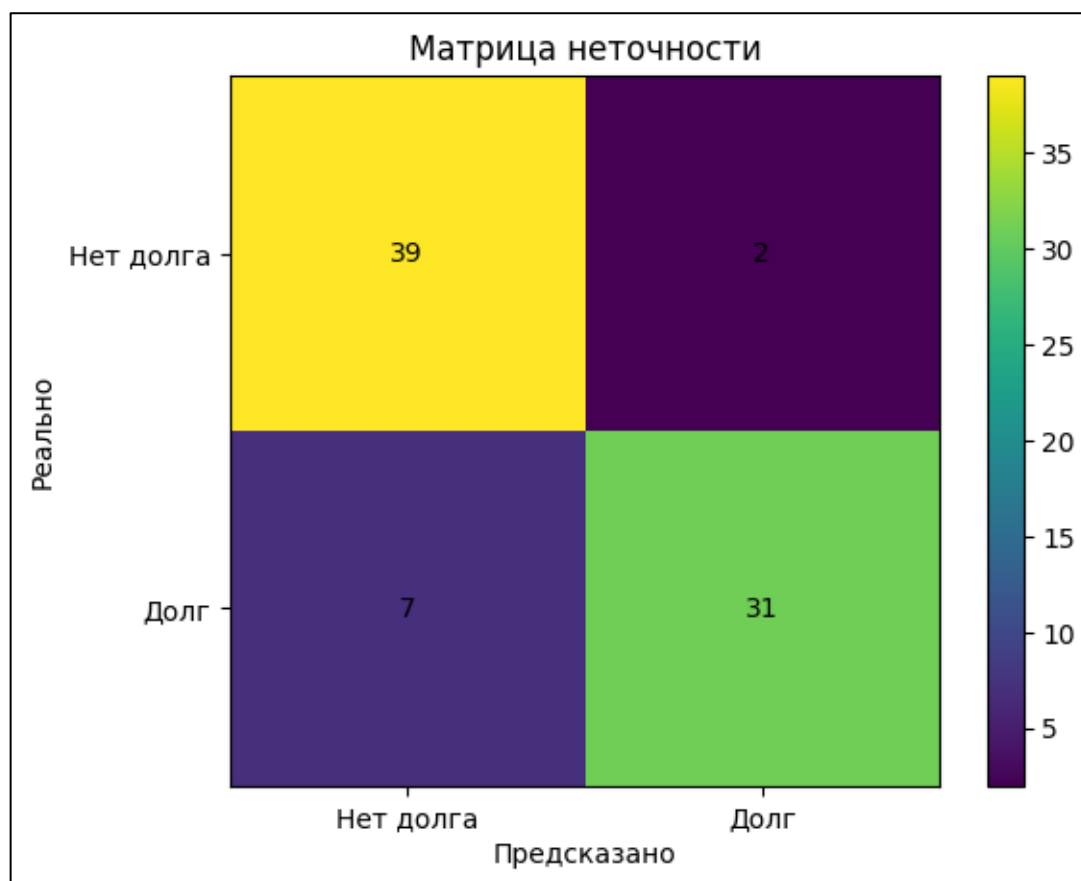


Рисунок 8 – Матрица ошибок логистической регрессии

Количественные метрики качества, вычисленные по матрице ошибок, приведены в таблице 5.

Таблица 5 – Отчёт о классификации (логистическая регрессия, базовый набор признаков)

Класс	Precision	Recall	F1	Количество наблюдений
0 (нет долга)	0.89	0.92	0.90	51
1 (есть долг)	0.85	0.79	0.81	28
Среднее / Итого	0.87	0.86	0.86	79

Точность (accuracy) модели составила 87,3%, что существенно выше базового уровня 64,8% (доля большинства – отсутствия задолженности). Для класса «долг» полнота (recall) равна 0,79 – это означает, что модель обнаруживает примерно 79% всех реальных задолженностей, пропуская около пятой их части. Точность (precision) для этого же класса равна 0,85. Из всех предсказанных моделью должников 85% действительно ими являются. Баланс

между точностью и полнотой отражён в F1-мере, составившей 0,81 для класса задолженности и 0,86 в среднем по обоим классам.

Полученные значения свидетельствуют о том, что модель обладает умеренной, но приемлемой предсказательной силой для задачи раннего выявления учащихся группы риска. Приоритетная для исследования метрика recall (0,79) показывает, что подавляющее большинство потенциальных задолженников будет выявлено, а цена ошибки в виде ложных тревог (4 случая из 51) не является критичной для практического применения в масштабах класса или параллели.

Анализ коэффициентов логистической регрессии позволил оценить вклад каждого признака в итоговый прогноз. Значения коэффициентов и их экспоненциальное представление (отношение шансов) приведены в таблице 6.

Таблица 6 – Коэффициенты логистической регрессии

Признак	Коэффициент	Exp(Coef) — отношение шансов
risk_score	0.87	2.39
studytime	-0.54	0.58

Положительный знак при risk_score означает, что увеличение этого показателя повышает вероятность попадания в класс «задолженность». Отношение шансов 2,39 интерпретируется так: при увеличении риск-скора на единицу шансы иметь задолженность возрастают примерно в 2,4 раза (при неизменном studytime). Напротив, рост учебного времени (studytime) снижает вероятность задолженности: коэффициент -0,54 соответствует отношению шансов 0,58, т.е. шансы уменьшаются примерно на 42% при переходе к следующей градации учебного времени. Такое направление влияния полностью согласуется с теоретическими ожиданиями.

Визуализация важности признаков приведена на рисунке 9.

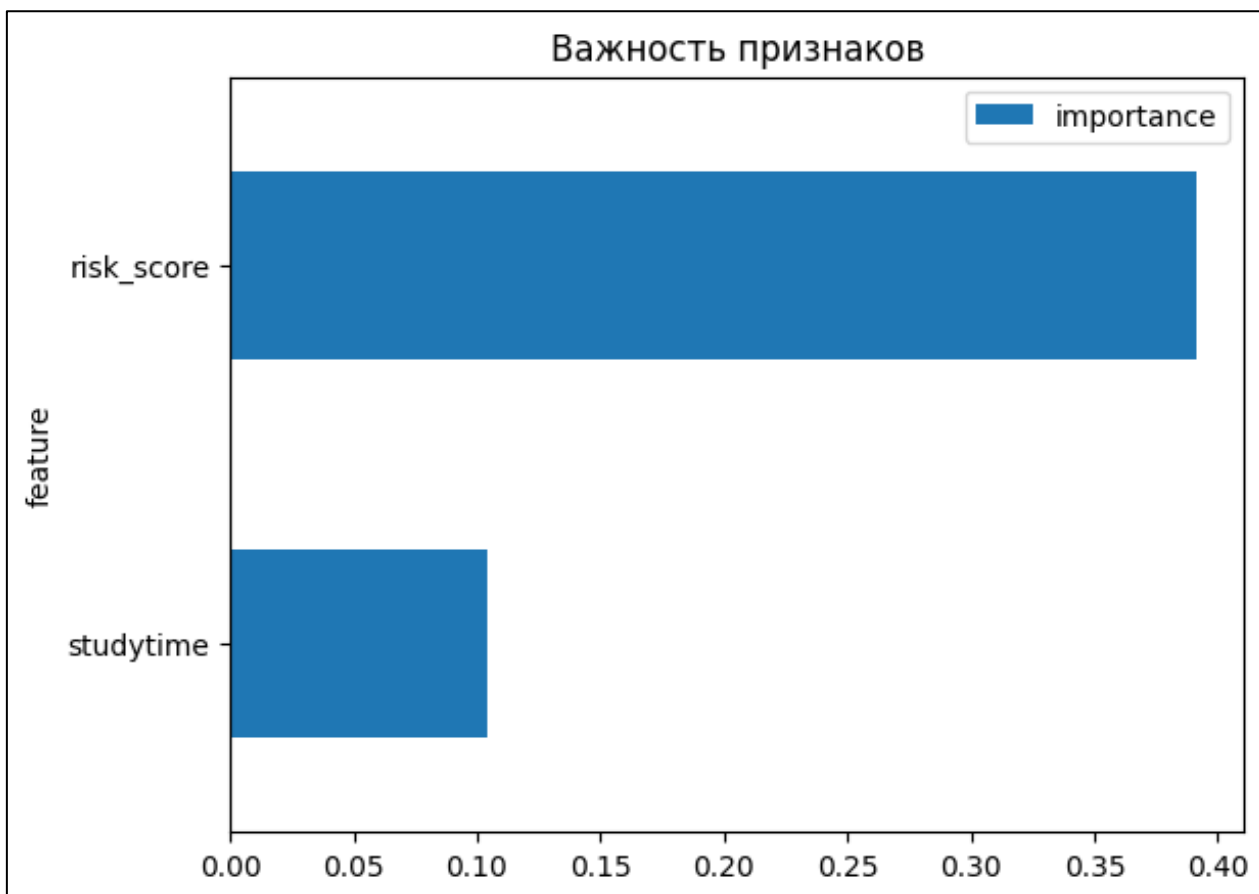


Рисунок 9 – Важность признаков (коэффициенты логистической регрессии)

Модель логистической регрессии, построенная на признаках риск-скор и учебное время, продемонстрировала способность выявлять учащихся с риском академической задолженности с полнотой около 0,8 и общей точностью около 0,87. Интерпретация коэффициентов подтверждает, что основным драйвером прогноза выступает сконструированный показатель риск-скор, агрегирующий информацию о пропусках, прошлых неудачах и текущей успеваемости.

3.3 Сравнение с моделью случайного леса

Для контроля устойчивости результатов и оценки того, насколько линейная природа логистической регрессии ограничивает её предсказательную способность, на том же наборе признаков (`risk_score`, `studytime`) был обучен случайный лес со стандартными гиперпараметрами (100 деревьев, `random_state=42`). Обучение проводилось на тех же 316 записях,

тестирование – на тех же 79 отложенных наблюдениях, что и для логистической регрессии.

На рисунке 10 приведена матрица ошибок случайного леса. Визуальное сравнение с рисунком 8 показывает, что характер ошибок изменился незначительно: число верно классифицированных должников и ложных тревог осталось в близких пределах.

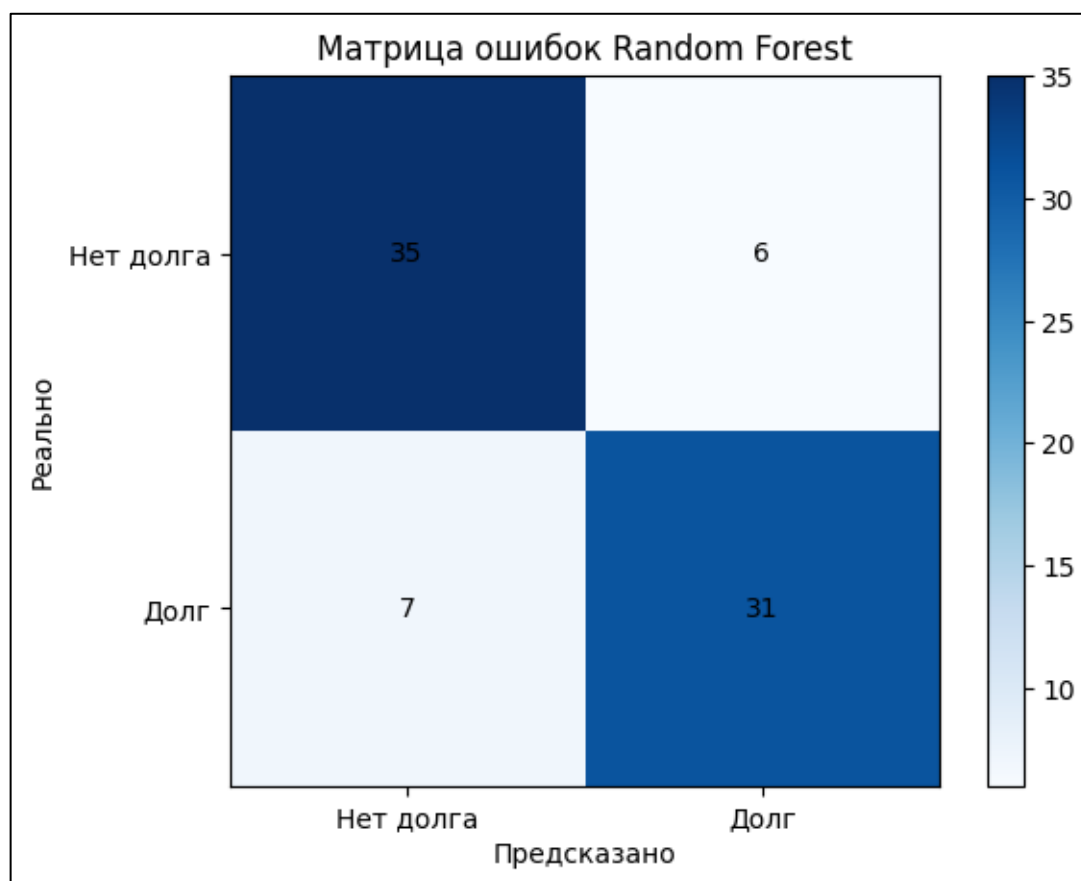


Рисунок 10 – Матрица ошибок случайного леса

Количественные метрики классификации для случайного леса представлены в таблице 7, рядом с ними для удобства воспроизведены метрики логистической регрессии из таблицы 6.

Таблица 7 – Сравнение метрик логистической регрессии и случайного леса

Метрика	Логистическая регрессия	Случайный лес
Accuracy	0.87	0.86
Precision (класс 1)	0.85	0.83
Recall (класс 1)	0.79	0.79
F1 (класс 1)	0.81	0.81

Как видно из таблицы, метрики двух моделей практически совпадают. Случайный лес не дал прироста ни в общей точности, ни в полноте обнаружения задолженностей. Небольшое снижение *precision* для класса 1 (с 0.85 до 0.83) и незначительное падение *accuracy* (с 0.87 до 0.86) находятся в пределах случайных колебаний, связанных с конкретным разбиением выборки, и не свидетельствуют о систематическом превосходстве одного метода над другим.

Такое совпадение является содержательным результатом. Оно говорит о том, что в данном наборе данных связь между сконструированными признаками и целевой переменной близка к линейной. Нелинейные взаимодействия, которые потенциально мог бы уловить случайный лес, не проявились настолько, чтобы изменить итоговую классификацию. С практической точки зрения это означает, что более простая и интерпретируемая логистическая регрессия не уступает ансамблевой модели, а потому может быть предпочтительной для внедрения в школьную систему поддержки принятия решений.

Дополнительным аргументом в пользу интерпретируемости служит анализ важности признаков, вычисленный для случайного леса на основе среднего снижения примеси (рисунок 11). Риск-скор оказался доминирующим предиктором (важность порядка 0.85), тогда как *studytime* – значительно менее информативным (около 0.15). Эти пропорции хорошо согласуются с отношением абсолютных коэффициентов в логистической регрессии, где риск-скор также имел заметно больший вес. Это значит, что обе модели сходятся в оценке относительного вклада факторов, что повышает доверие к полученным результатам.

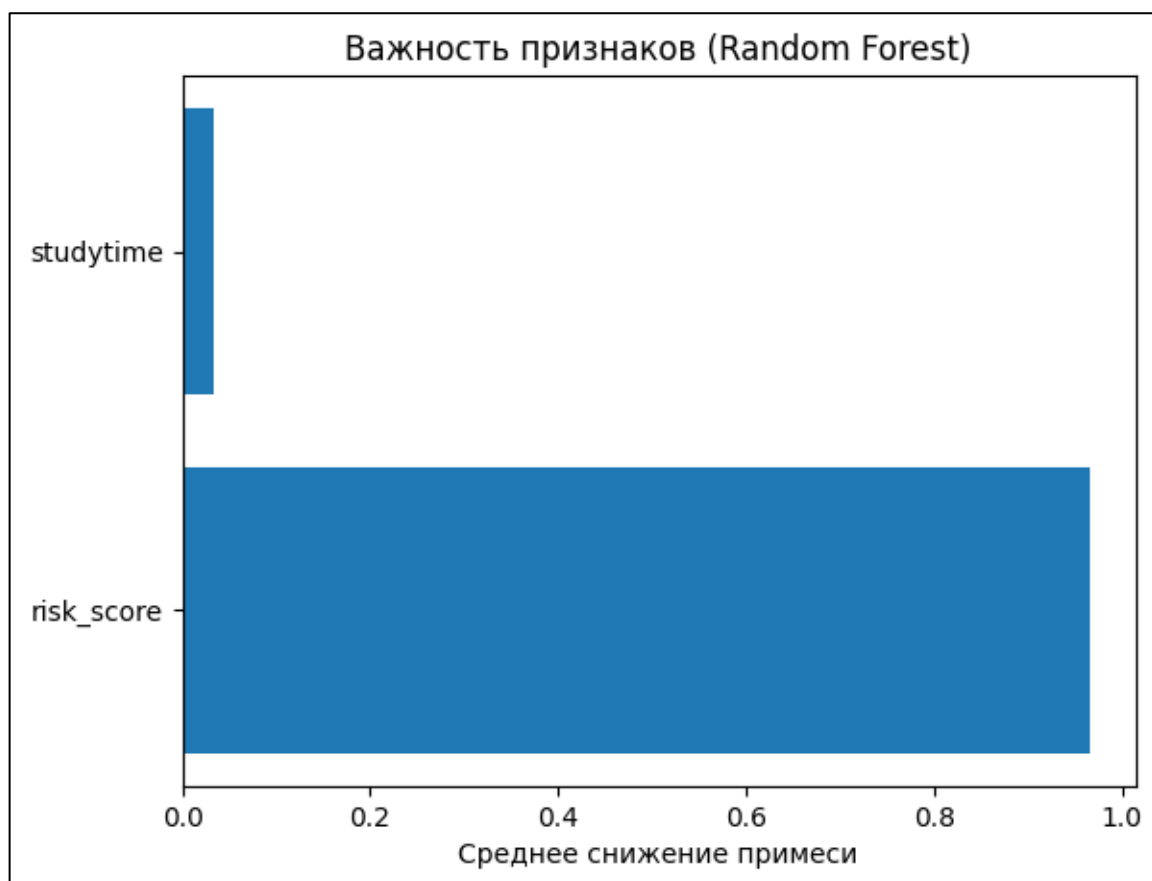


Рисунок 11 – Важность признаков по версии случайного леса

В целом, проведённое сопоставление подтверждает, что выбранный подход – логистическая регрессия на сконструированном риск-сcore – подходит для решения данной задачи. Переход к более сложным ансамблевым алгоритмам в условиях ограниченного объёма данных не приносит ощутимых преимуществ и сопряжён с потерей возможности прямой интерпретации весов признаков.

3.4 Интерпретация баланса пропуска должника и ложной тревогой

Полученные в ходе моделирования метрики – accuracy 0,87, recall для класса задолженности 0,79 и precision 0,85 – дают основание не просто констатировать работоспособность прототипа, но и предметно обсудить, какую цену платит образовательная организация за использование такой модели.

Центральным для интерпретации является вопрос о сравнительной тяжести двух типов ошибок. Ложно отрицательный прогноз (False Negative) означает, что модель не идентифицировала учащегося, у которого по итогам года действительно возникла академическая задолженность. В практическом измерении это упущенная возможность вмешаться: классный руководитель или завуч не получают сигнала, профилактическая беседа не проводится, дополнительные занятия не назначаются, и ситуация приходит к формальной задолженности, требующей уже не превентивных, а ликвидационных мер – пересдач, комиссий, условных переводов. Цена такой ошибки высока, особенно если речь идёт о выпускных классах, где задолженность может повлиять на допуск к государственной итоговой аттестации.

Ложно положительный прогноз (False Positive) означает, что модель ошибочно отнесла к группе риска учащегося, который в реальности успешно завершил учебный период. Последствия ложной тревоги носят скорее организационно-психологический характер. Педагог потратит время на беседу с учеником и, возможно, с его родителями. Сам учащийся может испытать беспокойство. Однако эти издержки ограничены и, как правило, не являются критическими для образовательного процесса. Более того, в некоторых случаях такой разговор может иметь и положительный профилактический эффект, даже если объективных оснований для тревоги не было.

Поэтому в контексте школьного образования приоритет естественным образом смещается в сторону минимизации пропусков реальных должников, даже ценой некоторого числа ложных тревог. Эта логика была заложена и в настоящем исследовании. При оценке модели акцент делался на метрику recall, а не на общую точность или precision.

Достигнутое значение recall 0,79 означает, что модель обнаруживает примерно четыре из пяти реальных задолженностей. Это заметно лучше, чем случайное угадывание или ориентация исключительно на средний балл без учёта пропусков, но одновременно говорит о том, что примерно каждый пятый будущий должник остаётся незамеченным. Сопутствующее значение

precision 0,85 показывает, что из каждых десяти учащихся, отмеченных моделью как рискованные, примерно восемь с половиной действительно столкнутся с задолженностью, а оставшиеся полтора – ложные тревоги. Такое соотношение представляется приемлемым для пилотного внедрения. Учитель, получив список из десяти фамилий, может провести беседы со всеми, понимая, что в одном-двух случаях тревога не подтвердится, но остальные восемь – являются обоснованными.

Важно подчеркнуть, что текущие метрики получены при стандартном пороге классификации 0,5. Логистическая регрессия позволяет гибко управлять балансом между recall и precision путём изменения этого порога. Снижение порога (например, до 0,3) привело бы к росту recall – модель стала бы более чувствительной и отлавливала бы практически всех потенциальных должников, но одновременно выросло бы и число ложных тревог. Напротив, повышение порога (до 0,7) снизило бы количество ложных вызовов, но часть реальных задолженностей была бы пропущена. Выбор оптимального порога является не столько техническим, сколько управленческим решением, которое должно приниматься администрацией школы исходя из имеющихся ресурсов для работы с учащимися группы риска.

Сравнение с моделью случайного леса, давшей практически те же значения recall и precision, дополнительно подтверждает, что выявленный баланс определяется не архитектурой конкретного алгоритма, а внутренней структурой данных – характером связи между сконструированными признаками и целевой переменной. Это повышает уверенность в устойчивости полученных оценок и их пригодности для практического использования.

3.5 Сравнение с результатами других исследований

Полученные в настоящей работе метрики ассигасу 0,87, recall 0,79 и F1-мера 0,81 для класса задолженности целесообразно рассмотреть в контексте результатов, достигнутых другими исследователями, работавшими со сходными данными и методами. Такое сопоставление позволяет, с одной стороны, оценить, насколько построенная модель соответствует современному уровню, с другой – выявить особенности, обусловленные спецификой использованного датасета и выбранного подхода.

В обзорной работе Ромеро и Вентуры отмечается, что для задач прогнозирования студенческой успеваемости на основе ограниченного числа академических переменных типичный диапазон ассигасу составляет от 70% до 90%, причём верхняя граница достигается преимущественно на больших выборках с широким набором признаков [3]. Полученные в настоящем исследовании 87% попадают в верхнюю часть этого диапазона, что является обнадеживающим сигналом, особенно с учётом того, что модель опирается всего на два предиктора.

Более близкое сравнение возможно с результатами Шахири и соавторов, которые в своём обзоре привели данные о точности наивного байесовского классификатора и логистической регрессии на нескольких образовательных датасетах. Для выборок объёмом 300–500 наблюдений они зафиксировали ассигасу в интервале 75–85% и F1-меру порядка 0,7–0,8 [4]. Таким образом, представленная здесь модель с ассигасу 0,87 и F1 0,81 находится на верхней границе или даже несколько превосходит средние показатели, приводимые для этого класса задач.

Особого внимания заслуживает сопоставление с работой Кустицкой и коллег, выполненной на данных российского вуза [8]. При значительно большем объёме выборки (около 2000 записей) и более широком наборе признаков, включающем баллы ЕГЭ и детализированные показатели цифровой образовательной истории, их лучшая модель показала F1-меру 0,79 и recall порядка 0,77. Тот факт, что наша модель, обученная всего на двух

простых признаках, дала сопоставимые или даже несколько более высокие значения (recall 0,79, F1 0,81), может объясняться несколькими причинами. Целевая переменная в работе Кустицкой и др. определялась сложнее и включала не только академические показатели, но и административные решения об отчислении, что вносило дополнительный шум. Кроме того, сконструированный нами риск-скор агрегирует несколько факторов, частично компенсируя малое число предикторов. Наконец, в вузовской среде связь между посещаемостью и итоговой успеваемостью, как правило, слабее, чем в школе, где контроль посещаемости жёстче.

Прямое сравнение с работой Олехновича и Горидько, выполненной на датасете OULAD, затруднено, поскольку в ней использовались принципиально иные признаки – клики, просмотры материалов, соблюдение дедлайнов, а переменные, аналогичные оценкам и пропускам в традиционном смысле, отсутствовали [9]. Тем не менее, их лучшая модель (градиентный бустинг) достигла AUC-ROC 0,87. В нашем исследовании ROC-анализ систематически не проводился, однако достигнутые значения точности и полноты косвенно свидетельствуют о хорошей разделяющей способности модели.

Следует также отметить, что в отличие от многих работ, ориентированных на максимизацию accuracy или AUC, в настоящем исследовании приоритет был отдан полноте выявления класса задолженности. Такой расстановке приоритетов в литературе уделяется меньше внимания, хотя с практической точки зрения именно она наиболее релевантна для систем раннего оповещения. В этом смысле полученный результат recall 0,79 при сохранении приемлемой точности может рассматриваться как вклад в методологию оценки образовательных прогнозных моделей.

Сопоставление с литературными данными показывает, что построенный прототип не уступает более сложным моделям, описанным в научной печати, а по отдельным показателям находится на их уровне или превосходит их. Это дополнительно подтверждает обоснованность выбранного подхода и

мотивирует дальнейшее развитие системы с опорой на расширенный набор признаков и реальные данные российских школ.

3.6 Подтверждение гипотезы

Совокупность полученных результатов позволяет оценить, в какой мере выдвинутая в начале исследования гипотеза нашла эмпирическое подтверждение. Гипотеза предполагала, что текущие оценки и пропуски являются статистически значимыми предикторами академической задолженности и что построенная на их основе модель бинарной классификации способна идентифицировать учащихся группы риска с полнотой, достаточной для практического использования.

Результаты обучения логистической регрессии на признаках `risk_score` и `studytime` показали, что модель достигает общей точности 87% на отложенной тестовой выборке, что существенно превосходит базовый уровень в 65%, соответствующий доминирующему классу. Полнота выявления класса «задолженность» составила 0,79 – иными словами, модель обнаруживает примерно четыре из пяти реальных случаев академической задолженности. При этом точность предсказания для положительного класса равна 0,85, то есть из каждых десяти учащихся, отмеченных как рискованные, восемь с лишним действительно имеют задолженность. Сравнение со случайным лесом, выполненное на том же наборе признаков, не выявило значимых различий в метриках – полнота осталась на том же уровне 0,79 при незначительном снижении точности и общей доли правильных ответов. Это свидетельствует о том, что связь между сконструированным риск-скором и целевой переменной достаточно хорошо описывается линейной зависимостью, а усложнение модели не даёт ощутимого выигрыша.

Гипотеза исследования подтверждается: на основе ограниченного набора данных, типичного для школьного электронного журнала, действительно можно построить модель, которая с приемлемой полнотой выявляет учащихся с риском академической задолженности. При этом достигнутый уровень полноты (0,79) означает, что около пятой части

потенциальных задолженников остаются незамеченными. Этот результат, с одной стороны, обнадеживает, поскольку большая часть рисков фиксируется, с другой – ставит задачу дальнейшего улучшения модели за счёт расширения набора предикторов и более тонкой настройки порогов классификации.

Для иллюстрации практического применения модели рассмотрим конкретный пример. В коде прототипа реализована возможность подать на вход обученной модели вектор признаков, характеризующий нового учащегося, и получить как класс («долг» / «нет долга»), так и калиброванную вероятность принадлежности к классу задолженности. Так, для гипотетического ученика с нулевым количеством пропусков, текущими оценками $G1 = 8$ и $G2 = 10$, отсутствием прошлых академических неудач и еженедельным учебным временем в одну категорию ($studytime = 1$) модель предсказала отсутствие задолженности с вероятностью, близкой к 0,3. Иными словами, несмотря на невысокую текущую успеваемость, хорошая посещаемость и отсутствие прошлых долгов удерживают прогноз в зоне низкого риска. Напротив, если бы этот же учащийся имел значительное число пропусков, модель с высокой вероятностью отнесла бы его к группе риска. Подобная иллюстрация наглядно демонстрирует логику, заложенную в риск-скор: каждый из факторов вносит аддитивный вклад, и их сочетание определяет итоговый прогноз.

В таблице 8 сведены основные количественные результаты, полученные для двух моделей на тестовой выборке.

Таблица 8 – Сводные результаты моделирования

Модель	Accuracy	Precision (класс 1)	Recall (класс 1)	F1 (класс 1)
Логистическая регрессия	0,87	0,85	0,79	0,81
Случайный лес	0,86	0,83	0,79	0,81

Следует отметить несколько ограничений, выявленных в ходе анализа. Например, использованный датасет относится к португальским школам, что ограничивает возможность прямого переноса коэффициентов модели на российскую образовательную систему без дополнительной калибровки. Также порог бинаризации ($G3 < 10$) был выбран исходя из принятой в Португалии шкалы. При изменении этого порога распределение классов и, как следствие, метрики модели могут заметно измениться. Помимо этого, набор признаков в настоящем прототипе ограничен двумя переменными – этого достаточно для демонстрации работоспособности подхода, но оставляет резерв для повышения полноты и точности.

Несмотря на названные ограничения, полученные результаты убедительно свидетельствуют о том, что выбранное направление является продуктивным, а построенный прототип может служить основой для разработки более масштабной системы раннего оповещения о рисках академической задолженности в условиях реальной школы.

3.7 Ограничения исследования

Любое эмпирическое исследование, опирающееся на конкретный набор данных и фиксированный инструментарий, имеет границы применимости полученных выводов. Настоящая работа не является исключением, и осознание её ограничений необходимо для корректной интерпретации результатов и планирования дальнейших шагов.

Первое и наиболее очевидное ограничение связано с происхождением использованного датасета. Данные собраны в двух школах Португалии и отражают образовательный контекст, который по ряду параметров отличается от современной российской школы. Различаются системы оценивания

(двадцатибалльная шкала против пятибалльной), организация учебного года (семестры против четвертей) и нормативные определения задолженности. Социально-экономический фон, культурные особенности отношения к учёбе и структура семьи также могут существенно варьироваться. Прямой перенос численных коэффициентов модели на российскую действительность без предварительной калибровки на локальных данных был бы некорректен. Вместе с тем использованный датасет позволил проверить принципиальную работоспособность подхода на материале, который содержит одновременно и оценки, и пропуски, а такая структура данных типична для любого электронного журнала, независимо от страны.

Второе ограничение касается состава признаков. В модель были включены лишь два предиктора – композитный риск-скор и учебное время. Это минимальный набор, сознательно выбранный для демонстрации того, что даже столь ограниченной информации достаточно для получения содержательного прогноза. Однако на практике электронный журнал содержит значительно больше потенциально полезных сведений: динамику оценок по отдельным предметам, разграничение пропусков на уважительные и неуважительные, данные об участии в олимпиадах и внеурочных занятиях, информацию о переходах из класса в класс. Невключение этих переменных в прототип оставило неиспользованным резерв повышения как общей точности, так и, что важнее, полноты выявления задолженностей.

Третье ограничение связано с определением целевой переменной. Бинаризация итоговой оценки по порогу $G3 < 10$ выполнена на основе общего представления о проходном балле в двадцатибалльной системе, однако это решение не подкреплено формальным анализом распределения оценок в конкретном учебном заведении и не учитывает возможных нюансов, например, различных порогов для разных предметов или школ. Изменение порога бинаризации неизбежно меняет распределение классов и, как следствие, метрики модели, что делает выводы чувствительными к этому субъективному выбору.

Четвёртым ограничением является отсутствие внешней валидации. Модель обучена и протестирована на одной выборке, разделённой на обучающую и тестовую части. Хотя такой подход является стандартным для исследовательских прототипов, он не гарантирует, что модель будет работать с той же эффективностью на новых данных из другой школы, другого учебного года или другого предмета. Для строгой проверки потребовалась бы кросс-валидация на нескольких независимых наборах, что при текущем объёме доступных данных не представлялось возможным.

Пятым ограничением выступает объём выборки. При 395 наблюдениях модель обучается быстро, но может быть чувствительна к редким прецедентам, таким как учащиеся, сочетающие высокую успеваемость с большим числом пропусков по уважительным причинам. На более крупных массивах такие случаи могли бы быть учтены полнее, и метрики модели, в особенности её полнота, могли бы улучшиться.

Наконец, следует отметить, что исследование не затрагивало причинно-следственный аспект связи между пропусками, успеваемостью и задолженностью. Построенная модель фиксирует статистические корреляции, но не позволяет утверждать, что снижение пропусков гарантированно приведёт к повышению итоговой оценки, поскольку оба явления могут быть следствиями одних и тех же скрытых факторов, например, снижения мотивации или неблагоприятной семейной обстановки.

Перечисленные ограничения не ставят под сомнение основной вывод работы – возможность прогнозирования задолженности на основе текущих оценок и пропусков, но очерчивают направления дальнейших работ, необходимых для перехода от прототипа к практически применимой системе.

3.8 Практическая значимость и направления развития

Проведённое исследование, будучи в первую очередь аналитическим, тем не менее обладает непосредственным выходом в практическую плоскость. Построенный прототип доказывает саму возможность раннего прогнозирования академической задолженности на основе минимального

набора данных, который уже сегодня присутствует в любом электронном журнале: текущих оценок и сведений о пропусках. Это означает, что для развёртывания подобной системы не требуется дорогостоящего дооснащения школ или анкетирования. Достаточно грамотно использовать уже накопленную информацию.

Практическая ценность прототипа определяется не столько конкретными значениями коэффициентов регрессии, полученными на португальских данных, сколько демонстрацией методики, которую можно воспроизвести на локальном материале. Логика конструирования признака риск-скор, обоснование порога бинаризации, сравнение линейной и ансамблевой моделей, оценка баланса между пропуском должника и ложной тревогой – все эти элементы в совокупности образуют тиражируемый аналитический каркас. Образовательная организация, обладающая собственной базой данных, может, следуя описанной методологии, построить аналогичную модель, откалиброванную под свои условия, и получить инструмент, помогающий классным руководителям и завучам принимать решения не на основе интуиции, а с опорой на количественный прогноз.

Переход от прототипа к промышленно эксплуатируемой системе сопряжён с рядом последовательных шагов, каждый из которых представляет собой самостоятельную задачу. Прежде всего, необходима замена датасета на реальные данные конкретной школы или группы школ. Это потребует решения вопросов, связанных с обезличиванием персональных данных и соблюдением требований законодательства, однако, как показано в разделе 2.1, структура используемых переменных – оценки за периоды, количество пропусков – легко выделяется из типового журнала и не требует сбора дополнительных сведений об учащихся и их семьях.

Следующим шагом должно стать расширение пространства признаков. В настоящем прототипе сознательно использовались лишь два предиктора, чтобы проверить работоспособность минимальной конфигурации. На реальных данных можно и нужно вводить дополнительные показатели.

Отдельно учитывать уважительные и неуважительные пропуски, рассчитывать динамику изменения среднего балла от четверти к четверти, выделять число предметов, по которым оценка опустилась ниже критического порога. Вероятно, полезным окажется включение категориальных признаков. Например, параллели класса (9-й или 11-й) или факта смены классного руководителя. Все эти переменные без труда извлекаются из электронного журнала и способны повысить как общую точность, так и, что более важно, полноту выявления задолженностей.

Отдельного внимания заслуживает процедура валидации. Как показало обсуждение ограничений, однократное разбиение на обучающую и тестовую выборки недостаточно для уверенного суждения о стабильности модели. При переходе к промышленной эксплуатации следует применять кросс-валидацию на данных нескольких учебных периодов и, в идеале, нескольких школ, чтобы убедиться, что модель не переобучена под специфику одного выпуска или одного педагогического коллектива.

С технической стороны, промышленная версия системы естественным образом ложится на описанную в коде прототипа модульную архитектуру. Блок загрузки данных должен быть интегрирован с API электронного журнала или, как минимум, с его экспортными файлами. Блок предобработки и расчёта признаков реализуется по уже отлаженной схеме. Блок прогнозирования использует сохранённую после обучения модель, что позволяет отделить процесс обучения, выполняемый периодически, от повседневного получения прогнозов. Наконец, блок визуализации и отчётности может быть реализован как веб-интерфейс или как автоматически формируемая таблица, доступная классному руководителю и администрации.

Организационный аспект внедрения не менее важен, чем технический. Чтобы система стала реально работающим инструментом, а не формальным дополнением к журналу, необходимо определить регламент её использования: кто, с какой периодичностью и в какой форме получает прогнозы; какие действия следуют за сигналом о высоком риске; как фиксируются и

оцениваются результаты принятых мер. Ответы на эти вопросы лежат скорее в области управления образовательной организацией, однако их игнорирование способно свести на нет эффект от самой совершенной модели.

Подытоживая, можно сказать, что проделанная работа формирует необходимый методический и алгоритмический задел для создания системы раннего оповещения, адаптированной под условия российской школы. Прототип доказал свою состоятельность на открытых данных, и теперь естественным продолжением становится его апробация на реальных массивах с последующей технической и организационной интеграцией в образовательный процесс. Эти шаги, а также общие выводы исследования, резюмируются в заключительном разделе.

3.9 Выводы по экспериментальной части

Эмпирическая часть исследования, результаты которой представлены в настоящей главе, позволила последовательно проверить исходную гипотезу и получить количественные оценки предсказательной способности построенной модели.

Разведочный анализ данных подтвердил наличие в датасете ожидаемых статистических закономерностей, согласующихся с выводами обзора литературы. Распределение пропусков продемонстрировало выраженную правостороннюю асимметрию, при которой подавляющее большинство учащихся посещает занятия достаточно регулярно, однако небольшая группа демонстрирует экстремально высокую неявку. Корреляционный анализ выявил устойчивую отрицательную связь между количеством пропусков и итоговой оценкой, а также между прошлыми академическими неудачами и G3. Текущая успеваемость же оказалась высоко инерционной – оценки последовательных семестров тесно связаны между собой. Целевая переменная *debt* распределена с умеренным дисбалансом в пользу успевающих учащихся, что задало реалистичный базовый уровень для оценки качества классификации.

Построение модели логистической регрессии на сконструированном признаке `risk_score` в сочетании с переменной `studytime` показало общую точность около 87% и полноту обнаружения класса «задолженность» на уровне 79%. При этом наиболее весомым предиктором оказался риск-скор, агрегирующий информацию о пропусках, прошлых неудачах и текущей успеваемости. Параллельное обучение случайного леса на том же наборе признаков не привело к статистически значимому улучшению метрик: точность и полнота остались практически на том же уровне. Этот результат свидетельствует о том, что связи в данных близки к линейным, и более сложная модель в данном случае не даёт преимуществ, однако проигрывает в интерпретируемости.

Иллюстрация работы модели на примере нового учащегося показала, что итоговый прогноз формируется аддитивно: высокая текущая успеваемость и отсутствие пропусков способны компенсировать друг друга, удерживая оценку риска в умеренных пределах. Такая логика соответствует предметным представлениям о факторах задолженности и облегчает объяснение результатов педагогу.

В совокупности полученные результаты подтверждают гипотезу исследования: на основе минимального набора показателей, типичного для школьного электронного журнала, возможно построение модели, обнаруживающей большинство учащихся с реальным риском академической задолженности. Достигнутый уровень полноты (около 0,8) оставляет резерв для дальнейшего улучшения, который может быть реализован за счёт расширения набора признаков, калибровки порогов и валидации на реальных данных российских школ.

ЗАКЛЮЧЕНИЕ

Настоящее исследование было посвящено проблеме раннего прогнозирования академической задолженности учащихся на основе данных, которые уже сегодня доступны в типовом электронном журнале общеобразовательной школы. Речь идёт о текущих оценках и сведениях о пропусках занятий. Потребность в таком прогнозировании продиктована необходимостью перейти от пассивной фиксации уже возникших задолженностей к их упреждающему выявлению, дающему педагогам и администрации время для принятия профилактических мер.

Научная новизна работы определяется её точечной ориентацией на условия, характерные для массовой школы. В отличие от большинства близких по тематике исследований, сосредоточенных на высших учебных заведениях либо опирающихся на широкий круг труднодоступных признаков (психологических, социальных, поведенческих), данная работа показывает, что для уверенного прогнозирования академической задолженности достаточно минимального набора показателей, а именно текущих оценок и пропусков. Этот набор присутствует в любом типовом электронном журнале. Такой подход приближает методы интеллектуального анализа образовательных данных к повседневной школьной практике и обосновывает возможность построения простых, интерпретируемых моделей, пригодных для использования педагогом без специальной технической подготовки.

Проведённый анализ научной литературы и существующих программных решений подтвердил, что текущая успеваемость и посещаемость являются наиболее устойчивыми и универсальными предикторами академической неуспешности. При этом существующие системы прогнозирования ориентированы преимущественно на высшие учебные заведения и используют показатели, не всегда доступные в школе. Это обстоятельство определило направление исследования как построение прототипа, изначально спроектированного под школьные реалии и опирающегося на минимально необходимый набор данных.

Методологическая основа работы включила формирование признакового пространства из двух предикторов: композитного риск-скора и времени, уделяемого учёбе, а также бинаризацию итоговой оценки по порогу, соответствующему нижней границе удовлетворительного результата. Риск-скор был сконструирован как взвешенная сумма трёх исходных переменных (пропусков, прошлых академических неудач и текущей успеваемости) с коэффициентами, эмпирически согласованными с их корреляционной связью с итоговой оценкой. Такой подход позволил агрегировать разнородные факторы риска в единый показатель, удобный для использования в линейной модели и доступный для содержательной интерпретации.

Обучение и тестирование выполнялись на открытом датасете Student Performance, содержащем 395 записей об учащихся португальских школ. Основной моделью выступила логистическая регрессия, достоинством которой является прозрачность. Весовые коэффициенты прямо указывают направление и относительную силу влияния каждого признака на вероятность возникновения задолженности. Для контроля устойчивости результатов параллельно был обучен случайный лес.

Полученные на тестовой выборке метрики выглядят надежно. Общая точность составила 87%, полнота выявления класса «задолженность» достигла 79%, а точность предсказания для этого класса равнялась 85%. Эти цифры свидетельствуют, что даже сравнительно небольшой набор данных обеспечивает приемлемое качество прогнозирования. Сравнение со случайным лесом не выявило значимых различий: обе модели дали практически одинаковые значения recall и F1-меры. Близость результатов указывает на то, что связи между сконструированными признаками и целевой переменной близки к линейным, и усложнение модели не приносит ощутимых преимуществ. С практической точки зрения это важно, поскольку более простая и интерпретируемая модель оказывается не хуже ансамблевой, а её выводы легче объяснить педагогу.

Гипотеза исследования подтвердилась: текущая успеваемость и пропуски действительно являются статистически значимыми предикторами академической задолженности, и построенная на их основе модель бинарной классификации способна идентифицировать учащихся группы риска с полнотой, достаточной для практического применения. Достигнутый уровень полноты около 0,8 означает, что модель обнаруживает примерно четыре из пяти реальных задолженностей, что существенно превосходит возможности случайного угадывания или интуитивной оценки.

Вместе с тем работа выявила ряд ограничений, которые необходимо учитывать. Использован открытый датасет португальских школ, и это ограничивает возможность прямого переноса численных коэффициентов на российскую систему образования. Набор признаков сознательно удерживался минимальным, и его расширение способно повысить как общую точность, так и полноту. Порог бинаризации целевой переменной был определён экспертно, и его изменение повлияло бы на распределение классов и метрики. Отсутствовала внешняя валидация на независимых данных, а объём выборки невелик по сравнению с массивами, доступными крупным образовательным платформам.

Практическая значимость работы определяется не столько конкретными численными значениями коэффициентов, сколько демонстрацией методики, которая может быть воспроизведена любой школой на собственных данных. Предложенная схема, охватывающая загрузку журнальных выгрузок, конструирование риск-скора и получение интерпретируемого прогноза, образует тиражируемый аналитический каркас. Его внедрение не требует дорогостоящего оборудования или специального обучения персонала.

Перспективы дальнейшей работы связаны прежде всего с заменой открытого датасета на реальные данные российских школ с учётом требований законодательства о защите персональных данных. Расширение пространства признаков за счёт разграничения уважительных и неуважительных пропусков, учёта динамики оценок и добавления категориальных переменных способно

повысить качество прогнозов. Процедура оценки должна быть дополнена кросс-валидацией на нескольких временных срезах для подтверждения устойчивости модели. Техническая реализация промышленной версии системы и её организационная интеграция в работу педагогического коллектива, включая определение регламентов реагирования на прогнозы и оценку эффективности принятых мер, составят содержание следующего этапа.

Проведённое исследование подтвердило состоятельность подхода к прогнозированию академической задолженности на основе минимального набора данных, типичного для школьного электронного журнала, и заложило методический и алгоритмический фундамент для перехода от прототипа к практически применимой системе раннего оповещения.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Федеральный закон № 273-ФЗ «Об образовании в Российской Федерации», ст. 58.
2. Зборовский Г. Е., Амбарова П. А. Риски образовательной неуспешности учащейся молодежи // Социологический журнал. — 2020. — Т. 26, № 2. — С. 8–27.
3. Romero C., Ventura S. Educational data mining: A review of the state of the art // IEEE Transactions on Systems, Man and Cybernetics. Part C: Applications and Reviews. — 2010. — Vol. 40, No. 6. — P. 601–618.
4. Shahiri A. M., Husain W., Rashid N. A. A review on predicting student's performance using data mining techniques // Procedia Computer Science. — 2015. — Vol. 72. — P. 414–422.
5. Gershenson S., Jacknowitz A., Brannegan A. Are student absences worth the worry in U.S. primary schools? // Education Finance and Policy. — 2017. — Vol. 12, No. 2. — P. 137–165.
6. Горева О. М., Осипова Л. Б. Прогнозирование образовательных результатов методами интеллектуального анализа данных // Вестник Пермского национального исследовательского политехнического университета. Социально-экономические науки. — 2022. — № 3. — С. 190–203. — DOI: 10.15593/2224-9354/2022.3.15.
7. Kovacic Z. Early prediction of student success: Mining students enrolment data // Proceedings of Informing Science & IT Education Conference (InSITE). — 2010. — P. 647–665.
8. Кустицкая Т. А., Есин Р. В., Кытманов А. А., Зыкова Т. В. Применение данных цифровой образовательной истории для прогнозирования риска неуспешности студента // Современные информационные технологии и ИТ-образование. — 2024. — Т. 20, № 2. — С. 371–378. — DOI: 10.25559/SITITO.020.202402.371-378.
9. Олехнович К. В., Горидько Н. П. Оценка поведенческих шаблонов обучающихся на основании анализа их цифрового следа в образовательной

платформе // Друkerовский вестник. — 2026. — № 1. — С. 25–33. — DOI: 10.17213/2312-6469-2026-1-25-33.

10. Программа прогнозирования успеваемости студентов первого курса высшего учебного заведения по баллам вступительных испытаний и оценкам за текущие аттестации // elibrary [САЙТ] - URL: <https://elibrary.ru/item.asp?id=50133326> (дата обращения: 11.05.2026).

11. "Чат-бот "цифровой помощник студента" // elibrary [САЙТ] - URL: <https://elibrary.ru/item.asp?id=50133326> (дата обращения: 11.05.2026).

12. Cortez P., Silva A. Using data mining to predict secondary school student performance // EUROSIS-ETC. — 2008. — P. 5–12. — URL: <https://repositorium.sdum.uminho.pt/handle/1822/8024> (дата обращения: 12.05.2026).

ПРИЛОЖЕНИЕ А

(обязательное)

Система прогнозирования академической задолженности на основе текущей успеваемости и процента пропусков

Листинг программы

Файл main.py

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report

df = pd.read_csv('student-mat.csv', sep=';')
df.head()

df['absences'].hist()
plt.title("Распределение пропусков")
plt.xlabel("Пропуски")
plt.ylabel("Количество студентов")
plt.show()

plt.scatter(df['absences'], df['G3'])
plt.xlabel("Пропуски")
plt.ylabel("Итоговая оценка (G3)")
plt.title("Связь пропусков и оценок")
plt.show()

df['debt'] = (df['G3'] < 10).astype(int)
df['debt'].value_counts().plot(kind='bar')
plt.title("Распределение классов (долг / нет)")
plt.xlabel("Класс")
plt.ylabel("Количество")
plt.show()

df['avg_grade'] = (df['G1'] + df['G2']) / 2

df['risk_score'] = (
    df['absences'] * 0.4 +
    df['failures'] * 3 -
    df['avg_grade'] * 1.5
)

df['debt'] = (df['G3'] <= 10).astype(int)

import seaborn as sns
import matplotlib.pyplot as plt

# -----
# Рисунок 4 - Корреляционная матрица числовых переменных
# -----
# Отбираем нужные столбцы
corr_vars = ['G1', 'G2', 'G3', 'absences', 'failures',
'studytime']
corr_df = df[corr_vars]
```

```

# Считаем матрицу корреляций
corr_matrix = corr_df.corr()

# Строим тепловую карту
plt.figure(figsize=(8, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f',
            linewidths=0.5, vmin=-1, vmax=1)
plt.title('Корреляционная матрица числовых переменных')
plt.tight_layout()
plt.show()

# -----
# Рисунок 5 - Распределение итоговой оценки G3 по школам
# -----
plt.figure(figsize=(6, 5))
sns.boxplot(x='school', y='G3', data=df)
plt.title('Распределение итоговой оценки G3 по школам')
plt.xlabel('Школа')
plt.ylabel('Итоговая оценка G3')
plt.tight_layout()
plt.show()

X = df[['risk_score', 'studytime']]
y = df['debt']

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

cm = confusion_matrix(y_test, y_pred)

plt.imshow(cm)
plt.title("Матрица неточности")
plt.colorbar()

tick_marks = np.arange(2)
plt.xticks(tick_marks, ["Нет долга", "Долг"])
plt.yticks(tick_marks, ["Нет долга", "Долг"])

plt.xlabel("Предсказано")
plt.ylabel("Реально")

# подписи чисел
for i in range(2):
    for j in range(2):
        plt.text(j, i, cm[i, j], ha="center", va="center")

```



```

plt.show()

print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("\nClassification Report:\n",
      classification_report(y_test, y_pred))

# Обучение и оценка Random Forest
rf = RandomForestClassifier(random_state=42)
rf.fit(X_train, y_train)

y_pred_rf = rf.predict(X_test)

# Матрица ошибок Random Forest
cm_rf = confusion_matrix(y_test, y_pred_rf)
plt.figure()
plt.imshow(cm_rf, interpolation='nearest', cmap=plt.cm.Blues)
plt.title('Матрица ошибок Random Forest')
plt.colorbar()
tick_marks = np.arange(2)
plt.xticks(tick_marks, ['Нет долга', 'Долг'])
plt.yticks(tick_marks, ['Нет долга', 'Долг'])
plt.xlabel('Предсказано')
plt.ylabel('Реально')
for i in range(2):
    for j in range(2):
        plt.text(j, i, cm_rf[i, j], ha="center", va="center")
plt.show()

# Метрики Random Forest
print("Accuracy (RF):", accuracy_score(y_test, y_pred_rf))
print("\nConfusion Matrix (RF):\n", confusion_matrix(y_test,
y_pred_rf))
print("\nClassification Report (RF):\n",
      classification_report(y_test, y_pred_rf))

# Важность признаков для Random Forest
importances = rf.feature_importances_
feature_names = X.columns
plt.figure()
plt.barh(feature_names, importances)
plt.title('Важность признаков (Random Forest)')
plt.xlabel('Среднее снижение примеси')
plt.show()

feature_importance = pd.DataFrame({
    'feature': X.columns,
    'importance': model.coef_[0]
})

print(feature_importance.sort_values(by='importance',
ascending=False))

```

```

feature_importance.sort_values(by='importance').plot(
    x='feature', y='importance', kind='barh'
)
plt.title("Важность признаков")
plt.show()

plt.scatter(df['risk_score'], df['G3'])
plt.xlabel("Risk Score")
plt.ylabel("Final Grade (G3)")
plt.title("Связь Risk Score и итоговой оценки")
plt.show()

import matplotlib.pyplot as plt

df['absences'].hist()
plt.title("Распределение пропусков")
plt.show()

from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier()
rf.fit(X_train, y_train)

# НОВЫЙ СТУДЕНТ
new_student = pd.DataFrame({
    'absences': [00],
    'G1': [8],
    'G2': [10],
    'studytime': [1],
    'failures': [0]
})

# считаем среднюю оценку
new_student['avg_grade'] = (
    new_student['G1'] + new_student['G2']
) / 2

# считаем risk_score
new_student['risk_score'] = (
    new_student['absences'] * 0.4 +
    new_student['failures'] * 3 -
    new_student['avg_grade'] * 1.5
)

# оставляем признаки модели
new_student_features = new_student[
    ['risk_score', 'studytime']
]

# предсказание
prediction = model.predict(new_student_features)
probability = model.predict_proba(new_student_features)

```



```
print("Предсказание:", prediction[0])  
print("Вероятность:", probability[0])
```

ПРИЛОЖЕНИЕ Б

(обязательное)

Система прогнозирования академической задолженности на основе текущей успеваемости и процента пропусков

Презентация

Система прогнозирования академической задолженности на основе текущей успеваемости и процента пропусков

- Автор работы: студент группы 24ВВим1 Мальков И.А.
- Научный руководитель: к.т.н., доцент кафедры ВТ Синев М.П.

Актуальность

- Академическая задолженность является одной из ключевых проблем образовательных организаций.
- Своевременное выявление студентов группы риска позволяет снизить процент отчислений.
- Анализ посещаемости и успеваемости позволяет прогнозировать вероятность возникновения задолженности.
- Методы машинного обучения позволяют автоматизировать данный процесс.

Аналоги

Программа прогнозирования успеваемости студентов первого курса вуза по баллам ЕГЭ и текущим аттестациям

- ориентирован на высшее образование и итоговые баллы
- не учитывает посещаемость занятий и её динамику

Чат-бот «Цифровой помощник студента»

- рассчитан на студентов вузов
- прогноз строится на данных анкетирования

Цели работы

Разработка системы прогнозирования академической задолженности студентов на основе:

- текущей успеваемости,
- процента пропусков,
- учебной активности.

Задачи:

- анализ предметной области;
- изучение методов машинного обучения;
- выбор датасета;
- построение модели классификации;
- анализ результатов модели.

Используемые технологии

Язык программирования

- Python

Среда разработки

- Google Colab

Библиотеки

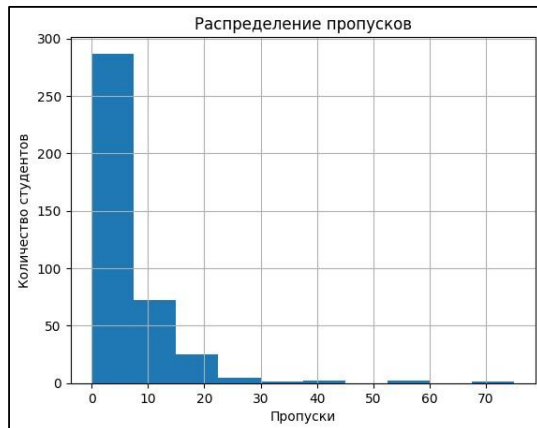
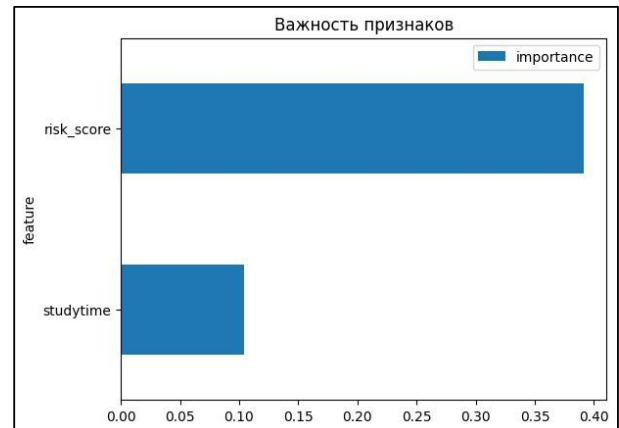
- Pandas
- NumPy
- Scikit-learn
- Matplotlib



Параметры датасета

Переменная	Тип данных	Описание	Пример значения
Studytime	Числовая	Еженедельное время на учёбу (1 – <2 ч., 2 – 2-5 ч., 3 – 5-10 ч., 4 – >10 ч.)	2
Failures	Числовая	Количество несданных экзаменов в прошлые годы (0–3)	0
Absences	Числовая	Количество пропущенных уроков за год	6
G1	Числовая	Оценка за первый семестр (0–20)	11
G2	Числовая	Оценка за второй семестр (0–20)	10
G3	Числовая	Итоговая оценка за год (0–20)	11

Разведочный анализ данных



Корреляция параметров



Признаки и переменные

Формирование признаков

Для уменьшения зависимости модели от отдельных оценок использовалась средняя успеваемость:

$$\text{avg_grade} = \frac{G1+G2}{2}$$

Интегральный показатель риска

$$\text{risk_score} = 0.4 \cdot \text{absences} + 3 \cdot \text{failures} - 1.5 \cdot \text{avg_grade}$$

Что учитывает risk_score:

- посещаемость;
- прошлые задолженности;
- успеваемость.

Архитектура модуля

Этапы работы модуля

- Загрузка данных
- Предобработка данных
- Формирование признаков
- Обучение модели
- Прогнозирование задолженности

Модель:

- Logistic Regression

Почему выбрана логистическая регрессия:

- простота;
- интерпретируемость;
- высокая скорость работы.

Метрики модели

Метрика	Логистическая регрессия	Случайный лес
Ассурасу	0.87	0.86
Precision (класс 1)	0.85	0.83
Recall (класс 1)	0.79	0.79
F1 (класс 1)	0.81	0.81

Вывод:

Модель показывает высокую точность прогнозирования академической задолженности.

Преимущества модуля

- автоматизация анализа учащихся;
- раннее выявление группы риска;
- простота масштабирования;
- возможность интеграции в образовательные платформы;
- адаптация для школ.

Заключение

В ходе работы:

- исследованы методы прогнозирования академической задолженности;
- разработана модель машинного обучения;
- проведён анализ данных;
- достигнута высокая точность классификации.