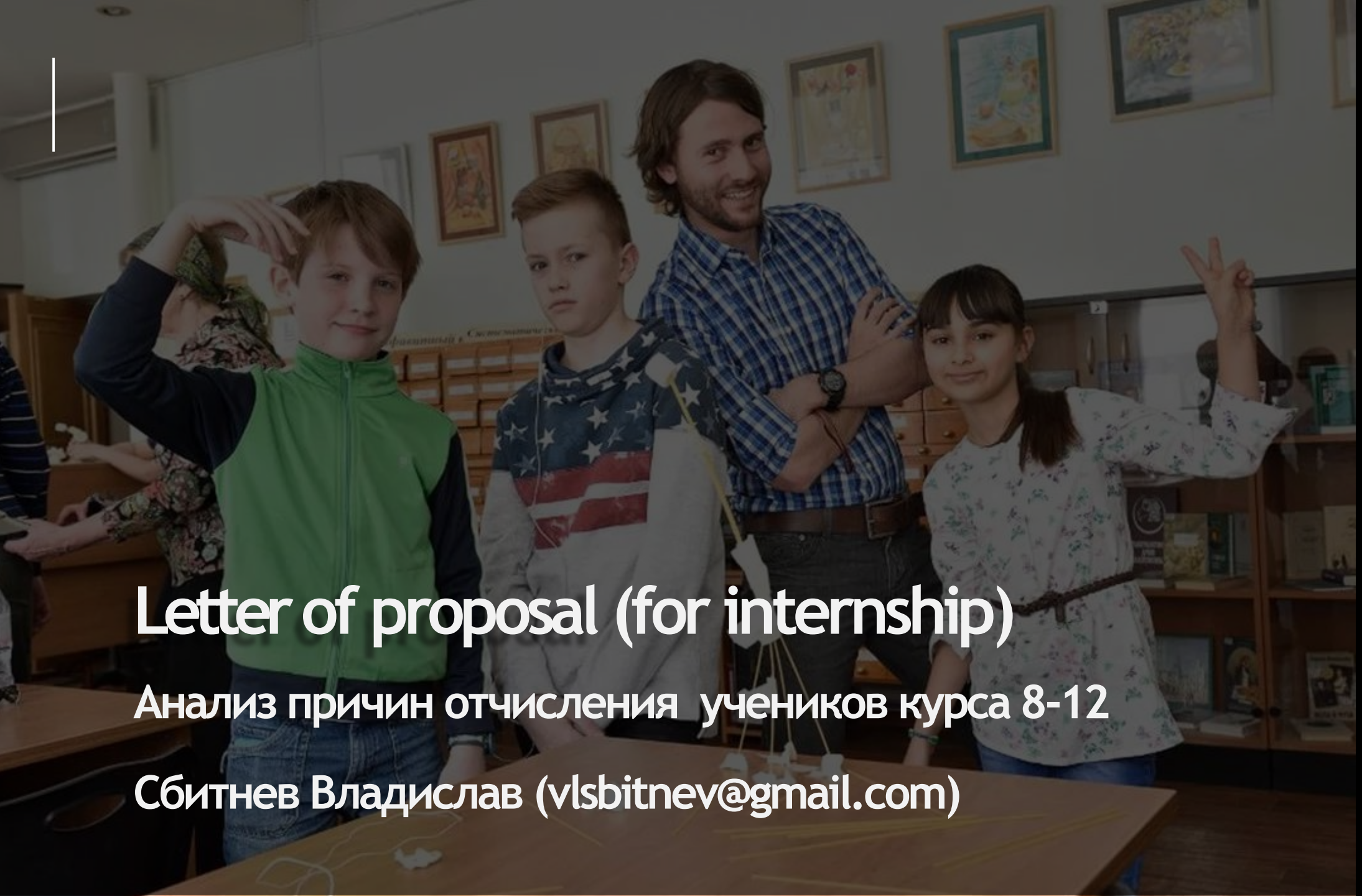


Letter of proposal (for internship)

Анализ причин отчисления учеников курса 8-12

Сбитнев Владислав (vlsbitnev@gmail.com)

Содержание

Executive Summary

Какие инсайды?	На что нам обратить внимание?	Какие рекомендации я бы мог дать департаментам компании?	Какие данные еще необходимы?
1. **Возраст** слабо влияет на вероятность быть отчисленным с курса 2. **Преподаватель** не влияет на вероятность ученика быть отчисленным 3. Проходившие больше **15 раз** скорее всего закончат обучение	1. Необходимо разработать **базовую предсказания модель отчисления по истории посещения занятий** для определения эффективности продуктовых и маркетинговых изменений в курсах 2. **Заполнение данных** – были пропуски в информации про учеников	1. Разработать методику «пробуждения» для а) ни разу не пришедших на занятия: при помощи триггер-емейлов/коммуникации с преподавателем б) чей паттерн посещение свидетельствует о том, что они перестают ходить на занятия	1. Собирать **больше соц-дем.иформации** о учениках для более точного моделирования учебного поведения 2. Данные по **другим курсам** для сравнения 3. **Большая выборка** учеников для уточнения модели

Источник: Проведенный анализ

Содержание

На основе предоставленных данных можно моделировать вероятность отчисления учеников в классе 8-12 в зависимости от прохождения курса

Предоставленный набор данных позволяет исследовать причины отчисления учеников с 8-12, а также моделировать вероятность учеников быть отчисленным в зависимости от его посещаемости

	Возраст	Дата старта	Дата последнего визита	Всего уроков	Прошло уроков	Педагог	Отчислен	Посетил уроков	Пропустил уроков	Дата отчисления
Тип переменной	Числовая	Дата	Дата	Числовая	Числовая	Категор.	Дата	Числовая	Числовая	Дата
Среднее	9.1 лет					Три педагога				

Для формирования репрезентативной выборки были отброшены студенты, не посещавшие курс 8-12 (3 строки)…

207320		04.03.19	21583	04.03.19	Николай	32	1	03.03.19	0	1	04.03.19	0		0	5-7
207319		04.03.19	21583	04.03.19	Николай	32	1	03.03.19	0	1	04.03.19	0		0	5-7
207009		03.03.19	21583	03.03.19	Николай	32	1	03.03.19	1	0	03.03.19	1		0	5-7

И студенты ни разу не посетившие занятия, так как у них отсутствует история посещений и на основе имеющихся данных невозможно объяснить их поведение (17 строк)

В модели необходимо использовать base rate (%), так как часть учеников всегда будет покидать курсы по различным неконтролируемым обстоятельствам

Источник: Предоставленный датасет

Необходимо вычислить дополнительные переменные, описывающие степень нагрузки на курсе, а также показывают на каком этапе находится группа на момент исследования

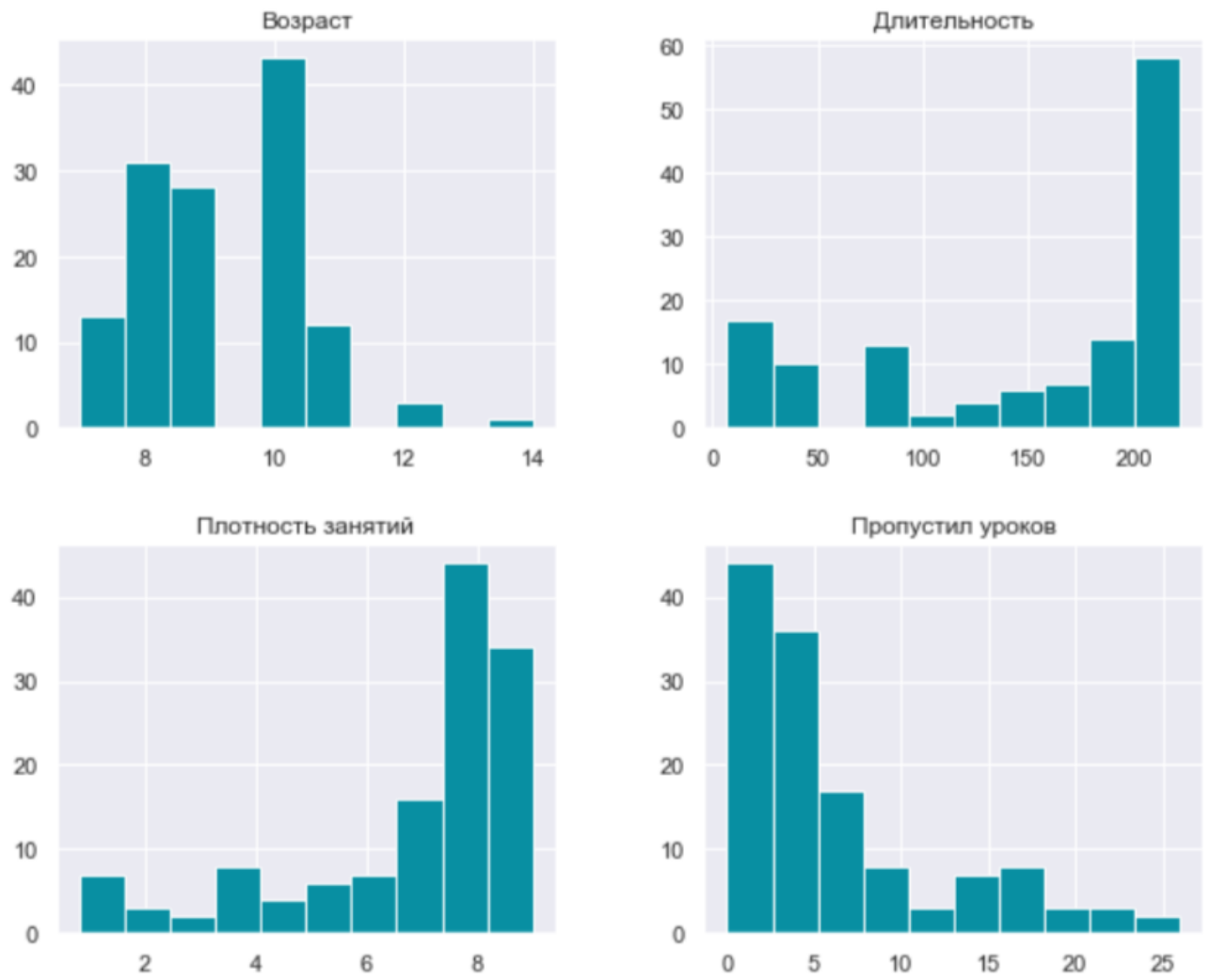
Переменные	Формула	Тип	Цель
Отчислен	= 1 – был отчислен, 0 – прошел до конца	Логический	**Зависимая переменная**
% прохождения	= Всего уроков ÷ Прошло уроков	Процент	На каком этапе прохождения сейчас находится группа
Длительность курса	= Последний визит - Старт группы	Числовой	Продолжительность занятий в днях
Частота занятий	= Длительность курса ÷ Старт группы	Числовой	Сколько в среднем проходит дней между занятиями

Источник: Предоставленный датасет

Содержание

Курс посещают ученики разного возраста, отчисление происходит как только ученик начинает прогуливать, нагрузка во всех группах в выборке одинаковая

Распределение ключевых числовых переменных



Выводы

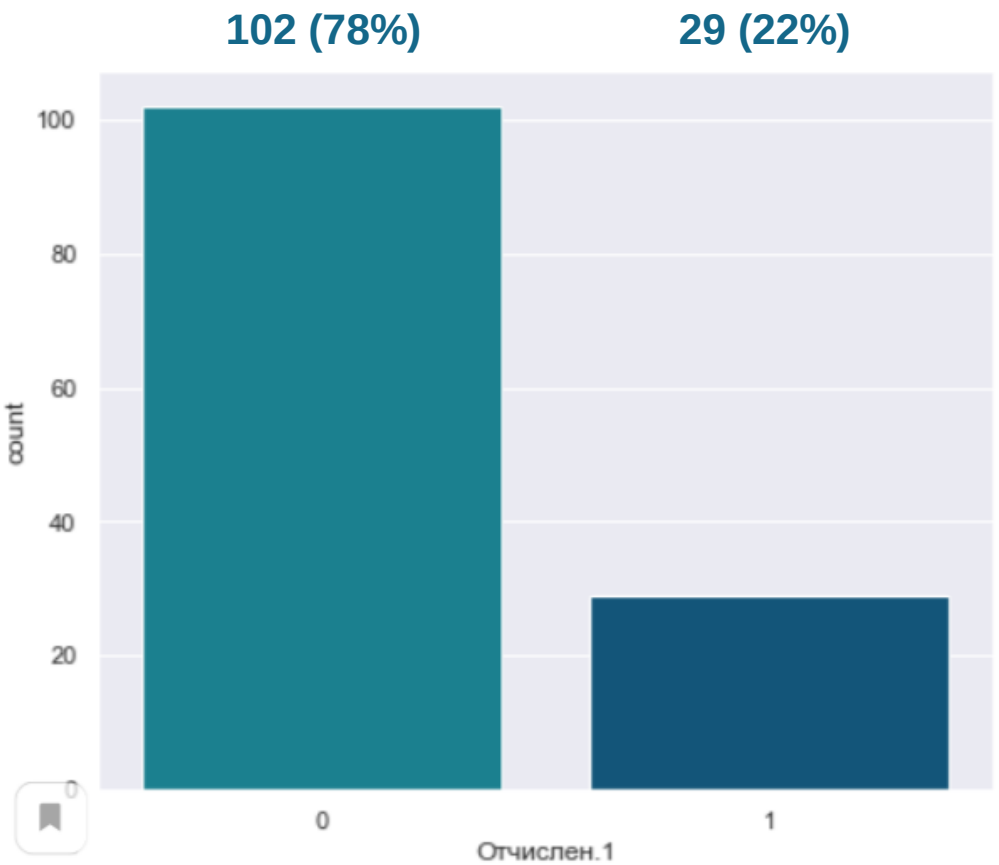
Из анализа распределения числовых переменных

- **Одинаковое число занятий в неделю.** Подавляющее большинство групп проводят занятия **1 раз в неделю**
- **Большой размах возраста.** Большой размах возраста может повлиять на вероятность быть отчисленным в силу особенностей раннего/позднего возраста
- **Отчисление чаще всего происходит как только ученик перестает посещать.** Характер числа пропусков свидетельствует об отсутствии «постоянных» прогульщиков – их быстро отчисляют

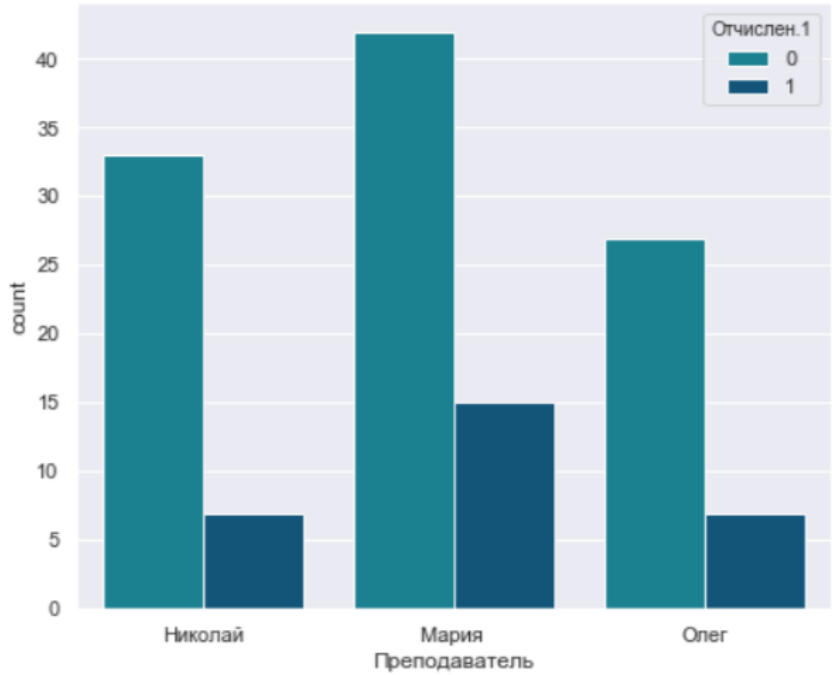
Источник: Предоставленный датасет

Общее число отчисленных студентов составляет 22%, при этом преподаватели в данной выборке не влияют значимо на вероятность студента быть отчисленным

Общий процент отчисленных составляет около 22% от всех учеников



Преподаватель не влияет значительно на вероятность студента быть отчисленным

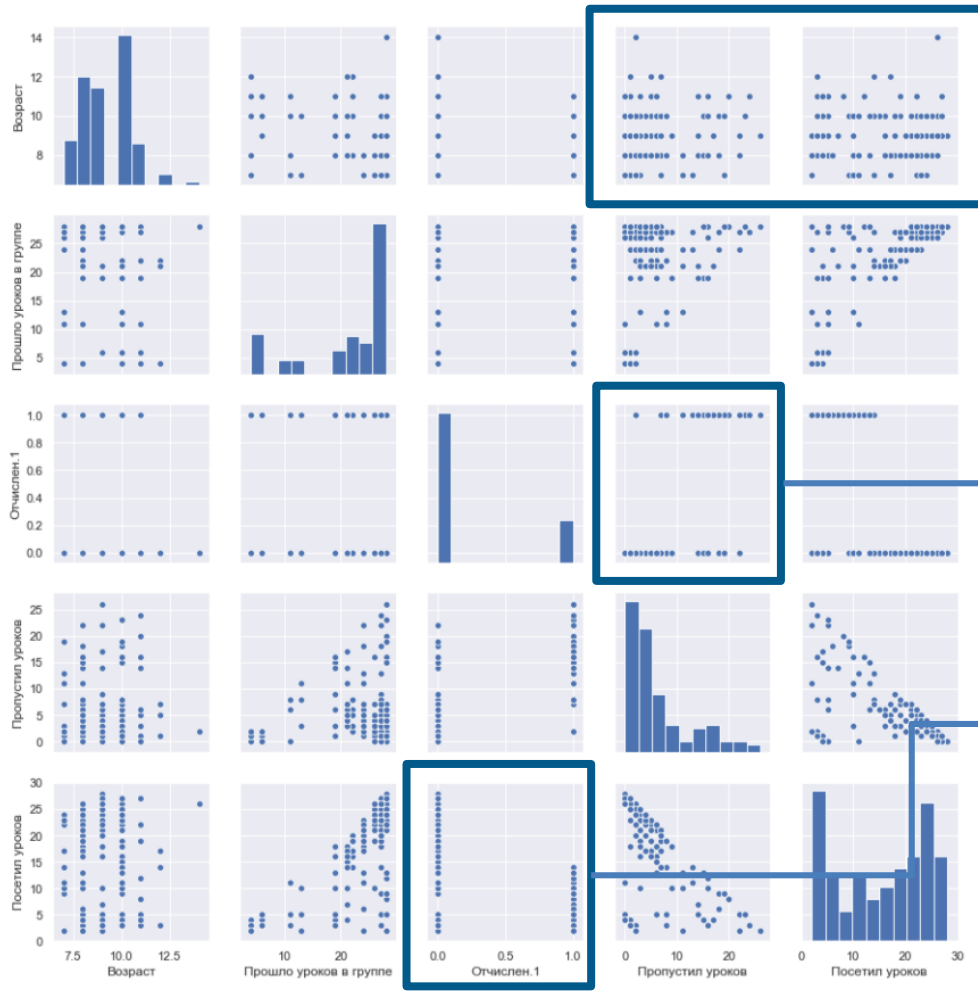


	Николай	Мария	Олег
Отчислен	42 (74%)	33 (82,5%)	27 (79%)
Обучается	15 (26%)	7 (17,5%)	7 (21%)
	100%	100%	100%

Источник: Предоставленный датасет

Необходимо подробнее изучить, как число посещенных занятий зависит от того, отчислят в итоге ученика или нет

Матрица взаимного распределения числовых переменных (по диагонали – распределение переменных)



Выводы

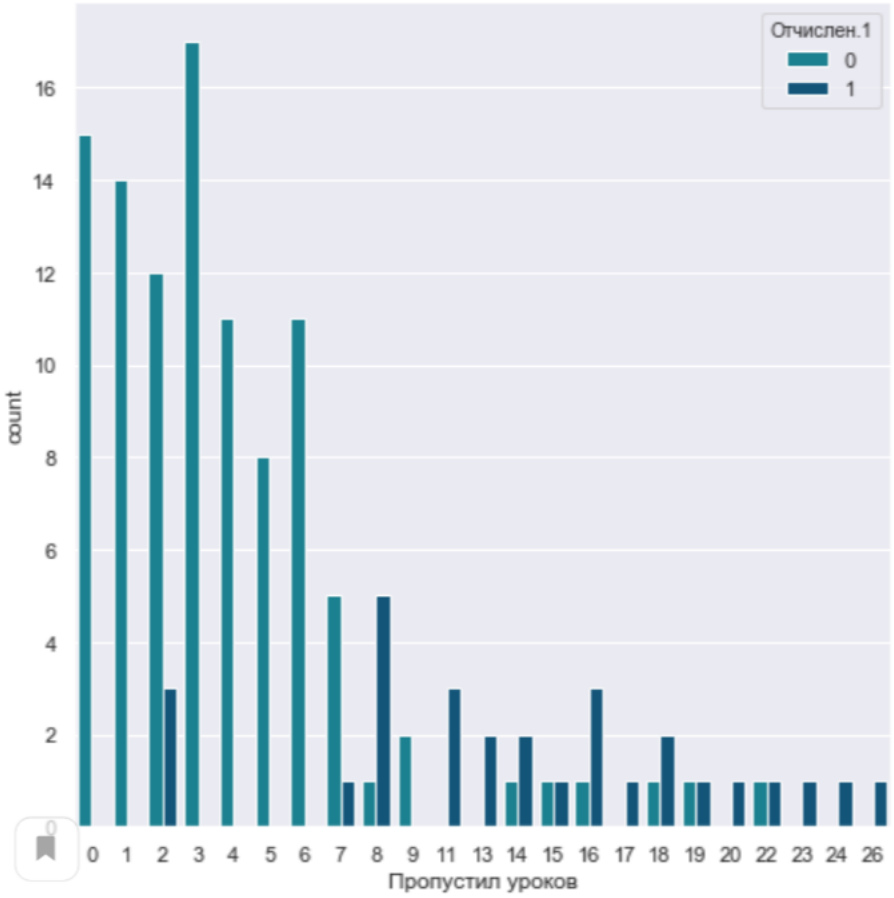
Из парного распределения переменных мы видим, что:

- **Отсутствует связь между возрастом и пропусками/посещением.** Посещение равномерно распределено в зависимости от возраста
- **Невозможно четко кластеризовать по пропущенным занятиям.** Причина – среди тех, кто успешно закончит также есть прогульщики. Более точным показателем следует признать показатель посещенных занятий.
- **НО! Отчисленные посещают меньше занятий.** После первых занятий, скорее всего, уже будет понятно, кого отчислят, а кого нет

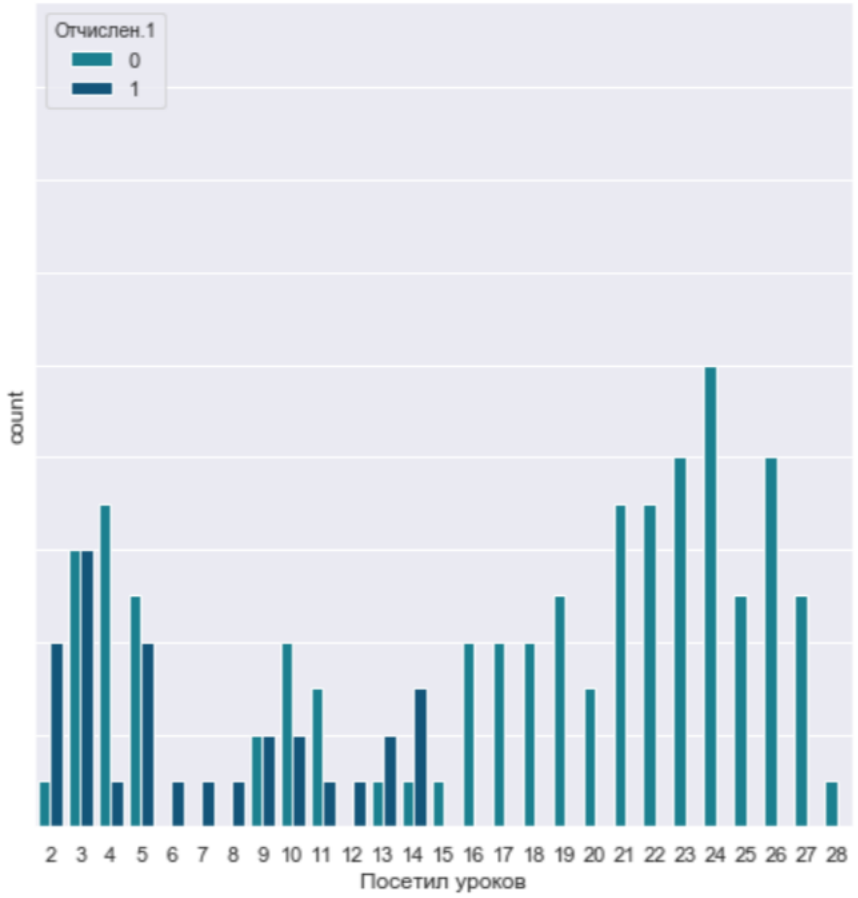
Источник: Предоставленный датасет

Ученики отчисляются на первых занятиях: осле 15 урока в данной выборке ученики заканчивали обучение

Отчисленные в среднем пропускают больше занятий



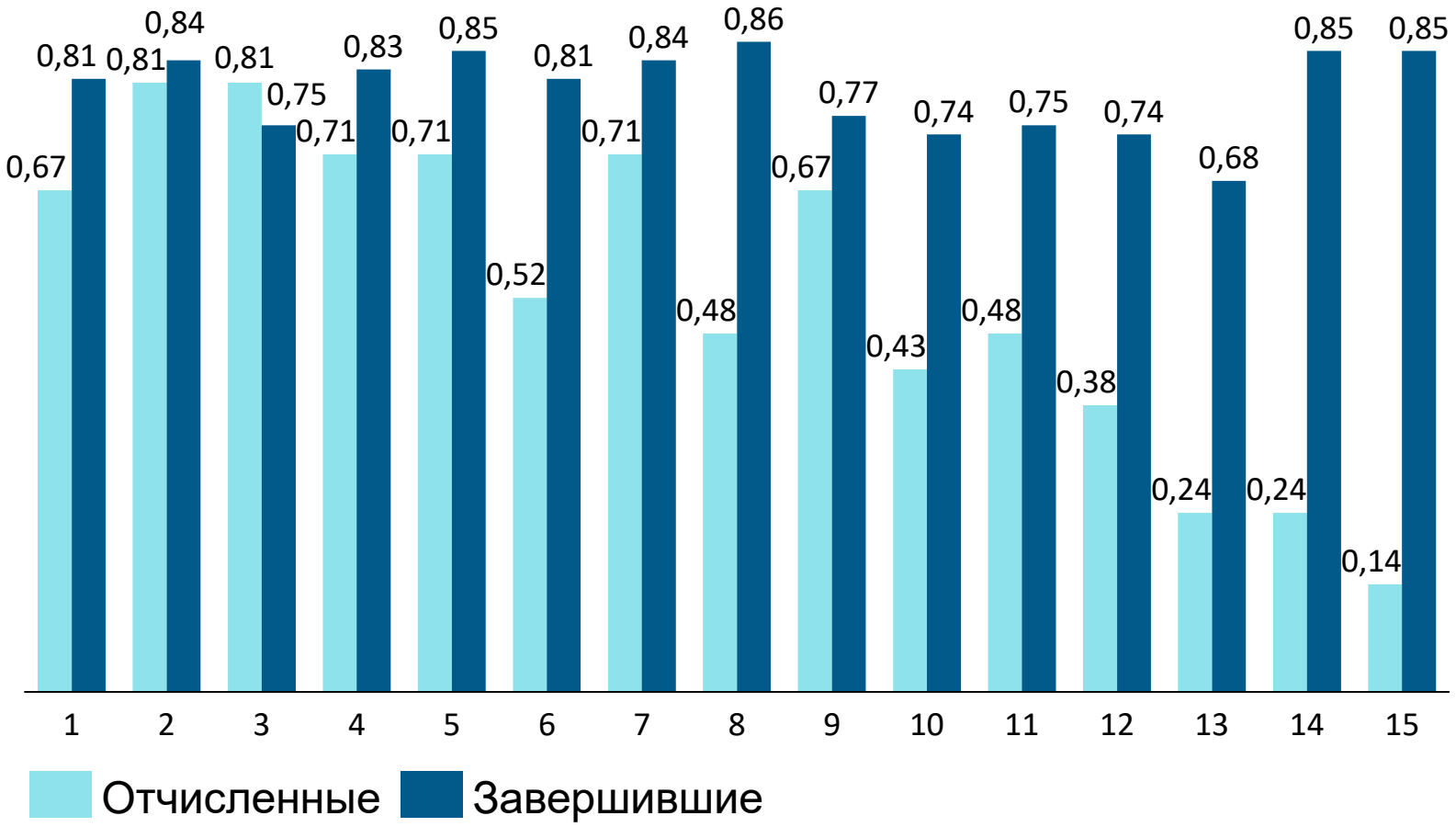
После 15 занятия ученики в данной выборке посещали курс до конца – все отчисления происходили в начале



Источник: Предоставленный датасет

Большая часть отчисляется в ходе первых 10 занятий, при этом на первых 5 занятиях отчисленные демонстрируют почти такую же посещаемость как успешно завершившие курс ученики

Вероятность посещения каждого из первых 15 курсов в зависимости от отчисления/завершения курсов



Расчет вероятности

- Взяты группы, завершившие обучение
- Вероятность посчитана как сумма пришедших на каждое занятие (отчисленные и завершившие отдельно), разделенное на общее число учеников в категории

Выводы

Статус ученика точно определяется после 9 занятия
Большая часть отчисляется до 10 занятия: можно явно различить две группы

Стоит отдельно проанализировать структуру курса, чтобы те, кто посетили первые 3-4 занятия остались заинтересованы

Источник: Предоставленный датасет

Содержание

На основе предоставленных данных можно моделировать вероятность отчисления учеников в классе 8-12 в зависимости от прохождения курса

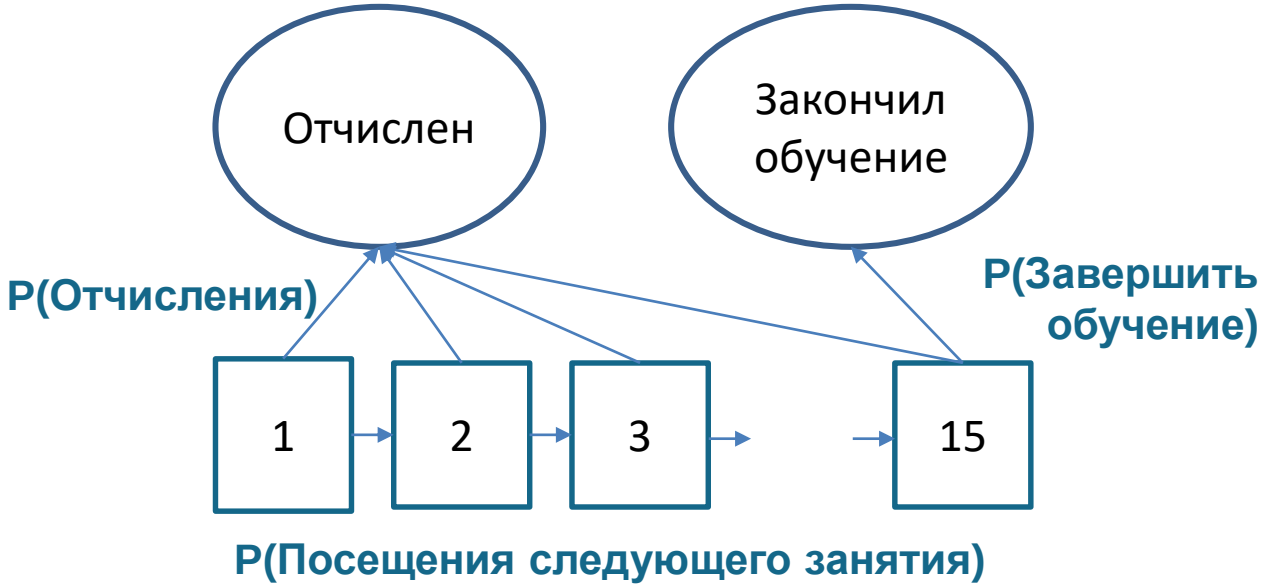
! Основная цель построения моделей – это не предсказание класса ученика, а моделирование текущего качества курса – при изменении программы/структуры курса можно будет количественно определять, насколько эффективными были внесенные изменения

Для решения похожих задач на практике используются три алгоритма

Logistic regression	Hidden Markov Chains	Naïve Bayes Classifier

Использовать «Посещаемость» как вектор бинарных объясняющих переменных

- Предсказание вероятности быть отчисленным по истории посещению каждого из 15 занятий



Определить «паттерны» посещения

- Наиболее часто встречающиеся последовательности в посещении занятий
- Вероятность отчисления по тому, как студент посещает занятия

Источник: https://arxiv.org/ftp/arxiv/papers/1802/1802.04845.pdfl https://stats.stackexchange.com/questions/322943/how-to-predict-employees-attendance-based-on-previous-attendance

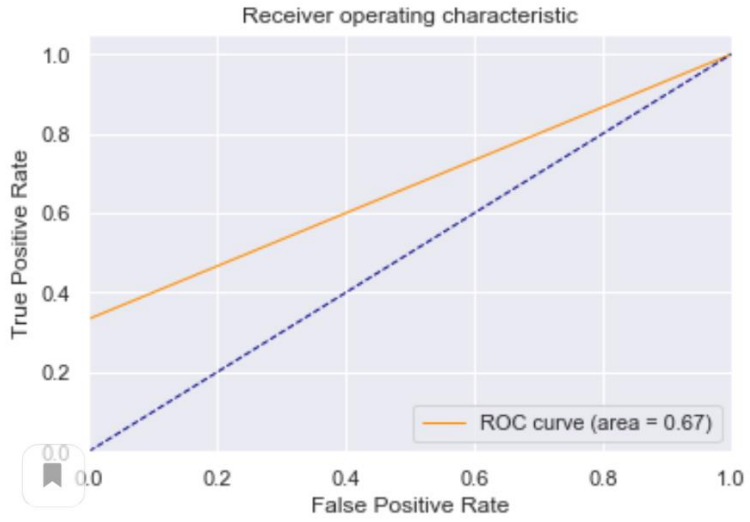
Содержание

Выборка слишком мала для создания значимой предиктивной модели

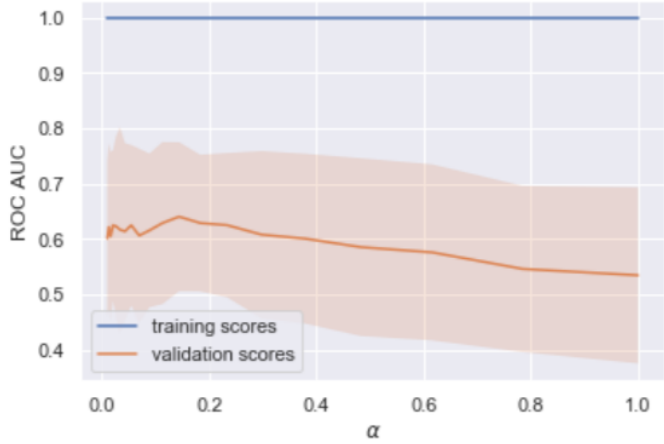
Для примера решения я проанализировал возможность использовать логистическую регрессию. Если мое решение покажется интересным, я готов в качестве стажера продолжить это исследование на более репрезентативной выборке

	Николай
Train-test split	0.2 по причине маленькой выборки
Модель	Логистическая регрессия

ROC-кривая



Performance модели в зависимости от выбора параметра регуляризации



Выводы

Небольшая выборка не позволяет модели набрать достаточное кол-во показателей: фактически она верно предсказывает 1 из 3 отчисленных из тестовой выборки и 18 из 18 продолживших обучение

Источник: Предоставленный датасет

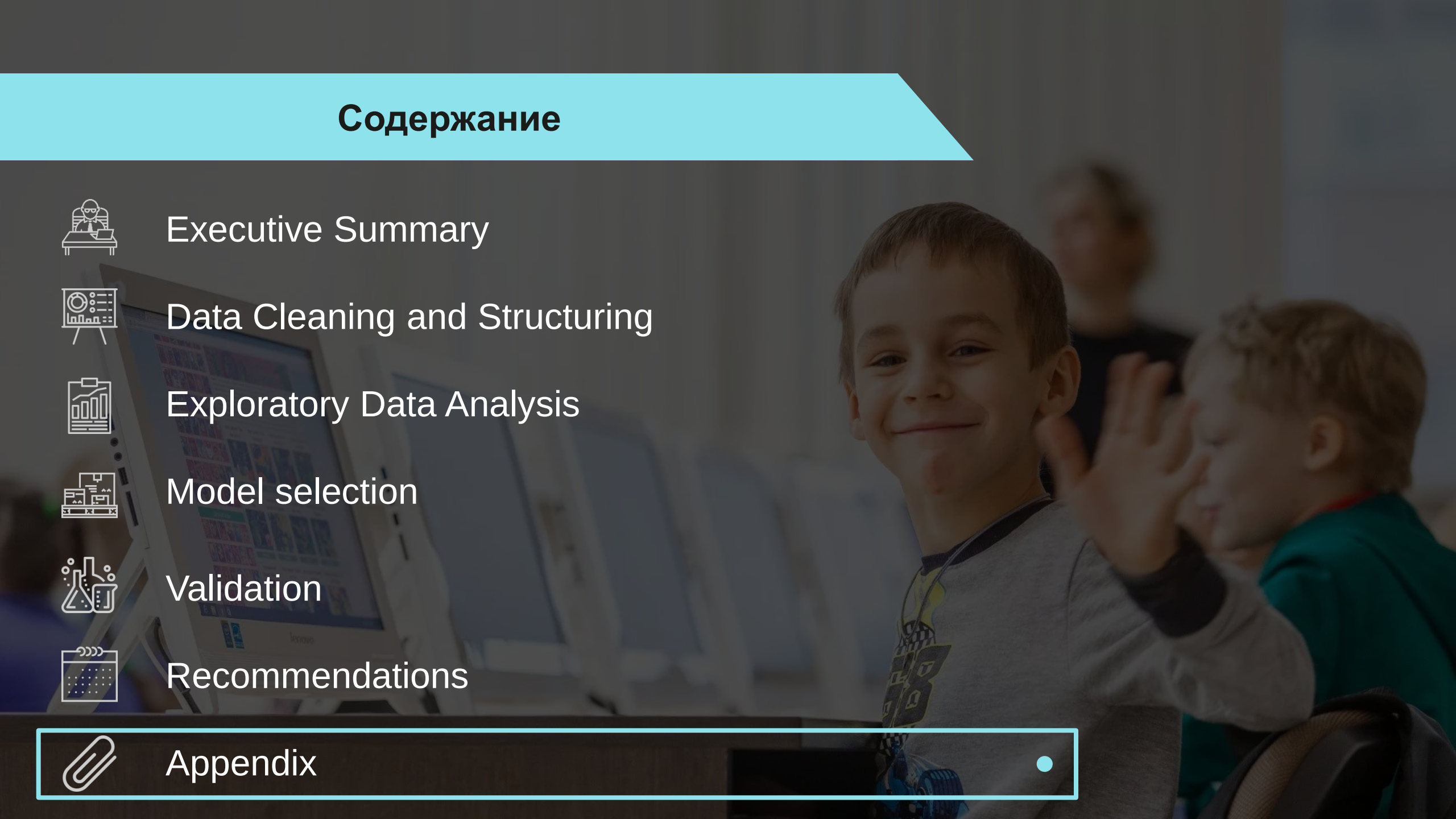
Содержание

Проведенный анализ показывает возможность моделирования поведения отчисленных учеников и определение вероятности их отчисления

Команда	Рекомендации
Маркетинговая команда	• Разработать методику «пробуждения» для: а) ни разу не пришедших на занятия: при помощи триггер-емейлов/коммуникации с преподавателем б) учеников, чей паттерн посещение свидетельствует о том, что они перестают ходить на занятия
Продуктовая команда **Методисты**	• Проанализировать структуру курса и состав заданий с точки зрения отчислений: изменить состав заданий, постараться увлечь на первых занятиях • Провести тесты: как меняется при изменении структуры курса % отчисленных
Преподаватели	• Собирать больше информации о учениках на платформе • Семинары и мастер-классы по тому, как заинтересовать детей и погрузиться в прохождение курса

Источник: Предоставленный датасет

Содержание



Запрос данных

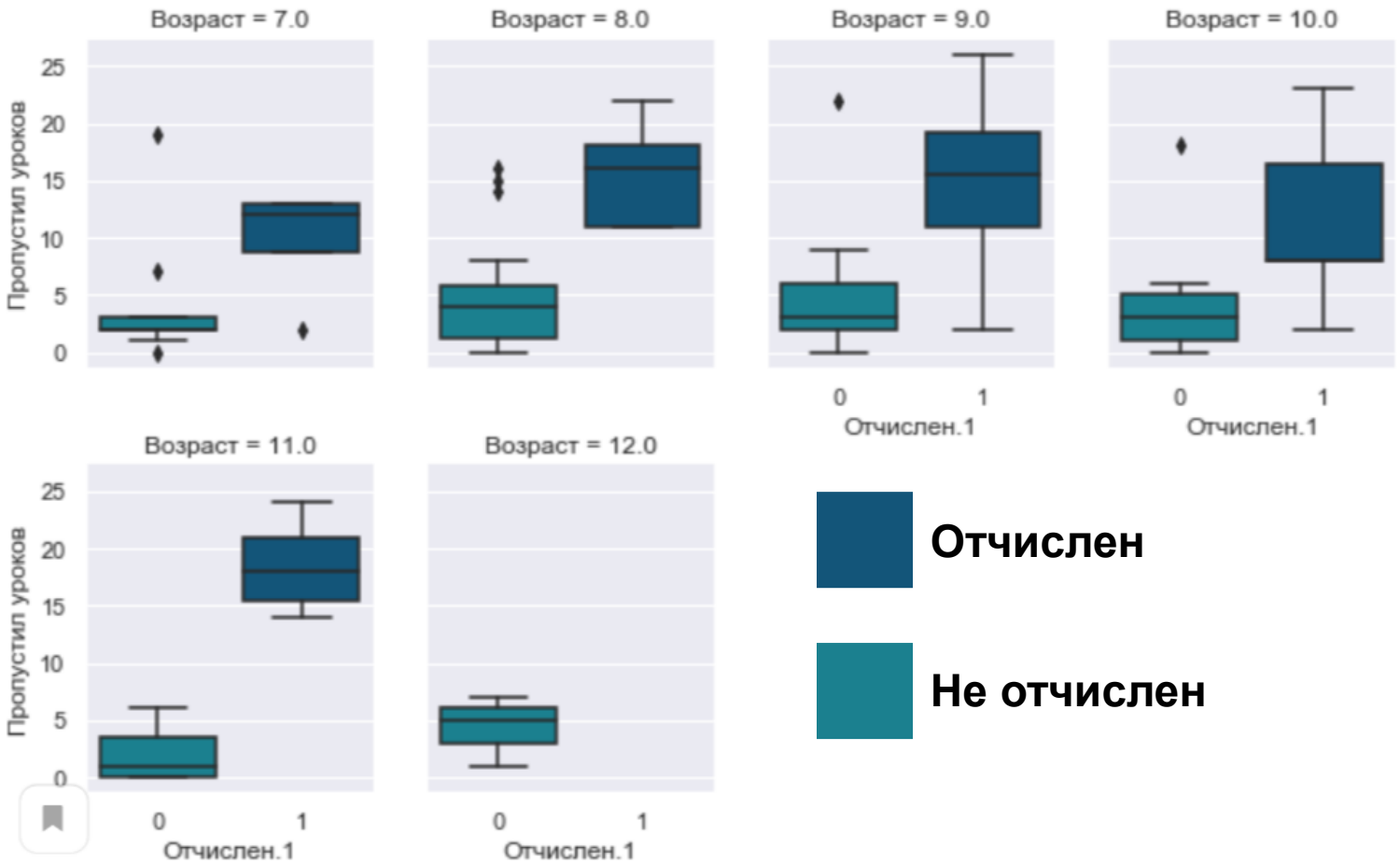
Для того, чтобы улучшить предсказательную точность модели нам необходимы следующие данные:

	Нам необходимы следующие данные…	…для того чтобы:
1	Больше социально-демографических характеристик наших учеников	Улучшить точность модели, провести более четкую сегментацию учеников по вероятности быть отчисленным
2	Более широкая выборка учеников	Улучшение точности модели, так как модель в работе обучена на незначительном объеме данных .которые могут нерепрезентативны

Источник: Анализ команды консультантов

Приложение 2. Визауализация паттерна пропуска уроков в зависимости от отчисления по возрастным когортам

Boxplot-ы пропусков в зависимости от возраста студента



Выводы

Возраст не влияет на паттерн пропусков учеников: во всех возрастных группах наблюдается похожая ситуация: те, кто был отчислен пропускали больше

Источник: Предоставленный датасет