**Таргет**

Таргетом (значением, которое необходимо предсказать) являлся средний балл студента за следующую сессию.

**Признаки для обучения**

Собраны все доступные признаки для обучения модели:

1. СКУД(посещение увниверситета)

* Среднее количество заходов в день в течение семестра,
* Среднее количество часов, проведенных внутри университета в день,
* Среднее время суток первого и последнего проходов в университет за семестр,
* Максимальное и минимальное количество минут и часов, проведенных в университете и за пределами университета в день за семестр (показывает, насколько долго отлучался студент между занятиями за пределы университета или, наоборот, как долго он был на занятиях),
* Суммарное количество минут и часов, проведенных внутри университета и снаружи (время отлучений между парами), за семестр,
* Среднее и медианное количество минут и часов, проведенных внутри университета и снаружи (время отлучений между парами), в день за семестр,
* Доля дней в семестре, когда студент посетил более 1, 2, 3, 4 пар
* Суммарное количество прикладываний пропуска за семестр,

Также были добавлены исторические данные по этим признакам, например, данные за предыдущий семестр или усреднение за несколько предыдущих семестров.

Эти признаки учитывают предполагаемую зависимость между посещаемостью университета, временем, проведенным внутри и успеваемостью студента.

1. Оценки за предыдущие семестры

Ниже перечисленные признаки были добавлены как за предыдущий семестр, так и взвешенные разными способами за несколько предыдущих семестров (EMA, SMA, EMA, WMA)

* Суммарный и средний балл за предыдущий семестр,
* Академическая группа

Эти признаки учитывают предполагаемую тенденцию успеваемости ученика за весь период обучения. Условно говоря, отличники на протяжении всей учебы так и показывают хорошие результаты.

1. Данные при поступлении

* Средний балл ЕГЭ
* Являлся ли студент призёром олимпиад
* Источник финансирования (бюджет, внебюджет, целевое обучение)

Сделано предположение, что поступившие по олимпиадам студенты, по высоким баллам ЕГЭ и остальные бюджетники будут иметь лучшую успеваемость, нежели студенты с полным возмещением затрат.

1. Средние оценки по предметам, основываясь на всех студентах

Что сделано:

Для каждого существующего предмета была найдена средняя оценка путем усреднения оценок всех доступных нам студентов.

Далее для каждого студента для каждой сессии собирался список предметов и:

* Находилась средняя, максимальная, минимальная оценки по этим предметам,
* Количество сдаваемых предметов на сессии

Эти признаки помогут учесть глобальную сложность предмета и общую тенденцию его сдачи.

1. Процент двоек

Ниже перечисленные признаки были добавлены как за предыдущий семестр, так и взвешенные разными способами за несколько предыдущих семестров (EMA, SMA, EMA, WMA)

* Процент оценок “неудовлетворительно”

1. Moodle

* Количество курсов, к которым приступил студент, за семестр,
* Приступил ли студент к курсу по английскому языку (курс английского языка есть у всех студентов вне зависимости от направления подготовки),
* Приступал ли студент к курсам вообще

Сделали предположение, что активное выполнение курсов на платформе Moodle коррелирует с успеваемостью

1. Общие данные о студенте

* Гражданство (Россия или иное),
* Пол,
* Имеет ли общежитие,
* Имеет ли льготу,
* Имеет ли студент какие-либо публикации (напр. научная статья)

**Модель**

Чтобы решить задачу регрессии предсказания среднего балла студента была выбрана модель градиентного бустинга Catboost. Catboost адаптирован под работу с категориальными признаками, которых в нашем датасете присутствует много. Также Catboost является оптимальной моделью по опыту предыдущих экспериментов. Также были испробованы модели линейной регрессии и еще один бустинг LightGBM, но были получены более низкие результаты.

**Результаты работы модели**

Получено удовлетворительное качество.

Метрики:

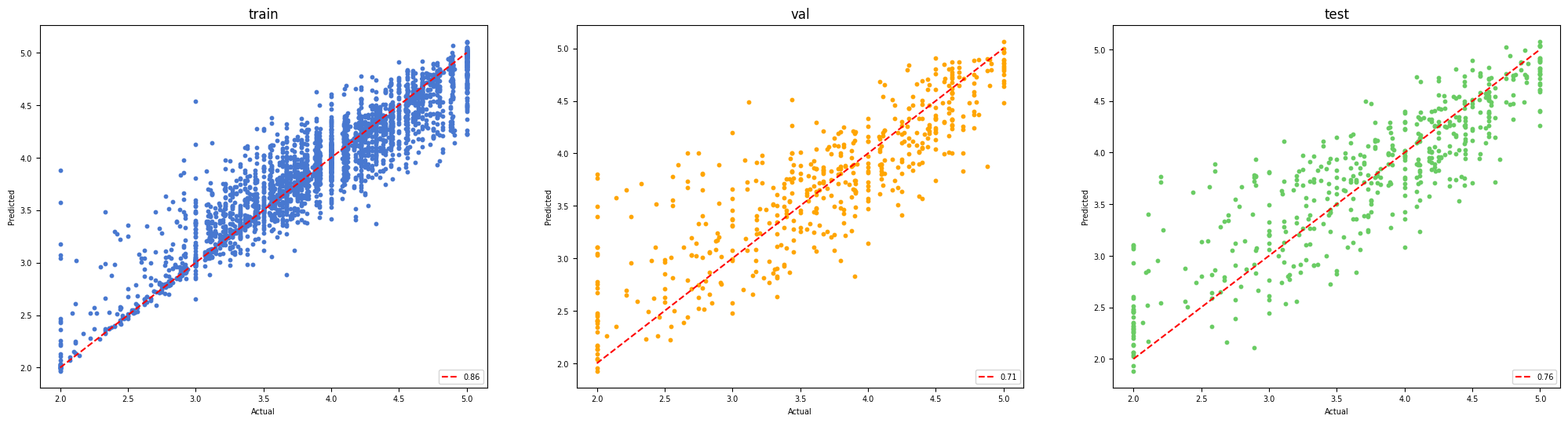
MSE: 0.16 – квадрат ошибки отклонения,

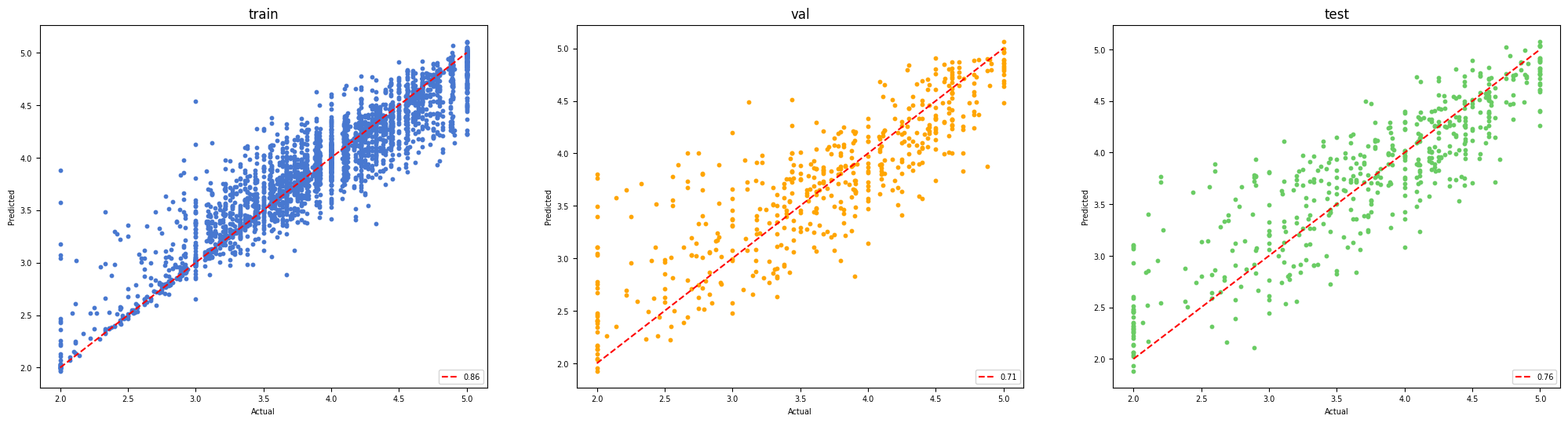
MAE: 0.31 – модуль ошибки отклонения. То есть в среднем, наше предсказание отличается от действительности на 0.31 балла.

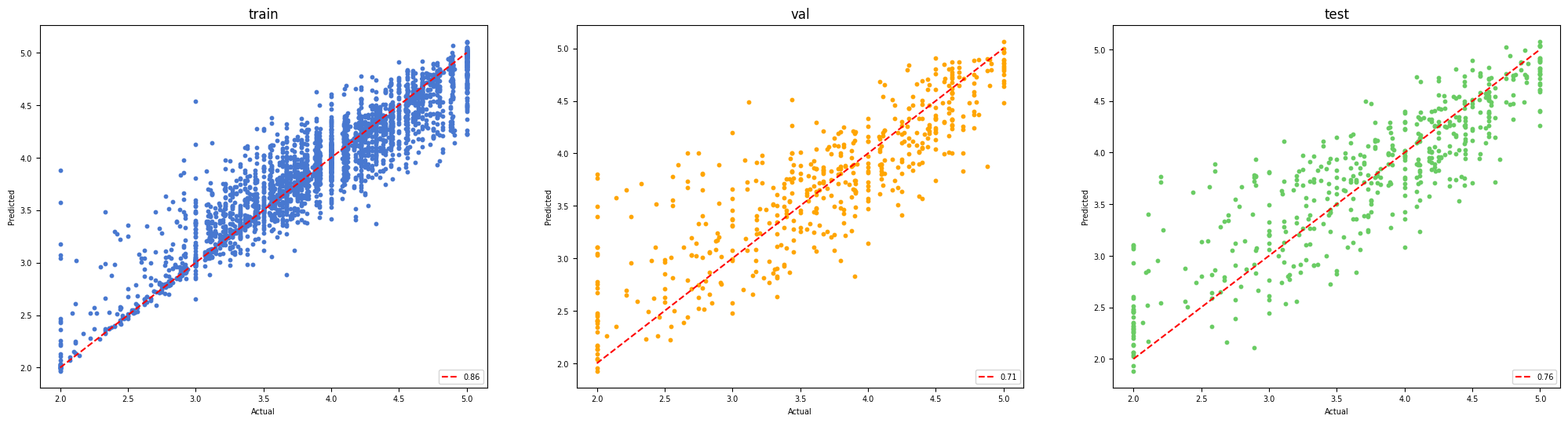
RMSE: 0.4 – корень из MSE

R2-score: 0.73 – главная метрика, на которую мы ориентировались. Показывает качество регрессии. Сравнивает дисперсии распределения таргетов, которые необходимо предсказать и таргетов, которые мы получили. Так как в нашей пятибальной системе разброс таргетов невелик, и ошибки MAE, MSE, RMSE могут быть искуственно занижены, поэтому R2-score является оптимальной метрикой. Показатель 0.73 является хорошим для этой метрики.

Судя по графикам, большую ошибку мы имеем, когда необходимо предсказать плохую оценку. На графиках ниже по горизонтали – оценка, которую надо предсказать, по вертикали – наше предсказание. Снизу, то есть там, где оригинальная оценка меньше, можно увидеть больший разброс как для тренировочной, так и для валидационной и тестовой выборок.



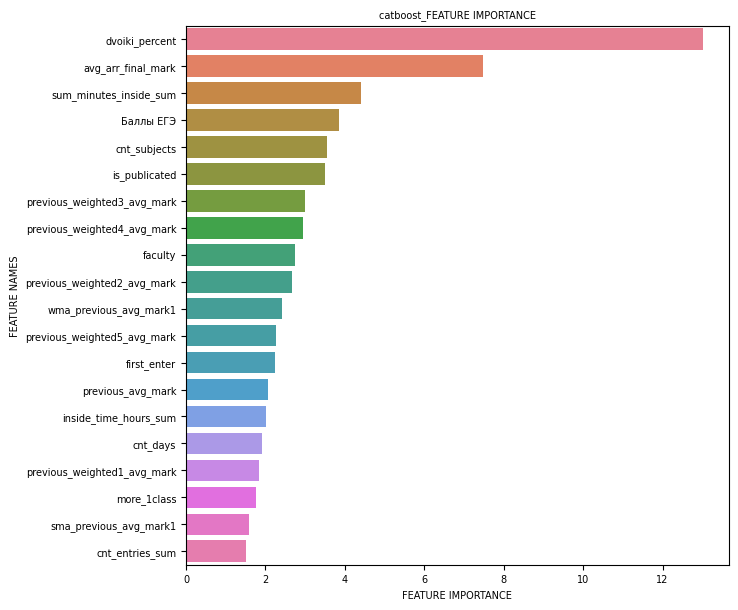




**Влияние признаков на модель.**

* Самым весомым признаком оказался процент двоек за предыдущую сессию.
* Далее идет усредненная оценка по предметам за сессию по всем студентам (подробнее в п. 4) – то есть есть влияние общей сложности предмета и тенденции его сдачи.
* Суммарное количество минут, проведенных в университете за семестр.
* Средний балл ЕГЭ.
* Количество сдаваемых предметов за сессию.
* Имеет ли студент какие-либо публикации (напр. научная статья).
* Оценки за предыдущие сессии.
* Направление подготовки.
* Усредненное время первого захода в университет.
* Количество дней за семестр, в которые студент посещал университет
* Доля дней за семестр, когда студент посетил одну и более пар.
* Суммарное количество прикладываний пропуска за семестр.

Вышеперечисленные признаки являются расшифровкой графика ниже, где по вертикали – признак, по горизонтали – его важность. Масштаб оси Feature Importance для нас не имеет значения, для нас важно лишь соотношение важностей признаков друг к другу. Например, признак процента двоек у студента за предыдущую сессию является набиолее важным признаком, который более, чем вдвое важнее остальных признаков.



**Дальнейшие улучшения модели**

Так как система электронных курсов Moodle была введена только с осеннего семестра 2023 года, мы не имеем предыдущих данных о выполнении. Ранее была система LMS Canvas, при возможности объединить данные о выполнении курсов и в Moodle и в Canvas, мы бы получили полную информацию о выполнении курсов и смогли бы проверить гипотезу влияния качества выполнения электронных заданий на успеваемость студента.

Тщательная проверка данных и очищение их от выбросов.