Работа с общим датасетом

1. Были скачаны оба датасета за 2023 и 2024 года и объединены записи в них (это можно сделать, так как, во-первых, структура датасетов была одинаковая, а во-вторых, ID студентов и ID персон в датасете за 2023 год не совпадали с ID студентов и ID персон в датасете за 2024)

2. Было замечено, что на каждый предмет у каждого студента приходится как минимум две строчки: у каждого предмета есть одновременно и значение оценки в ведомости, и итоговая оценка. Для того, чтобы избавиться от дубликатов, были удалены все строки, в которых колонка "UUID ведомости" была пустая (то есть были оставлены оценки из ведомости, а не итоговые).

3. Далее датасет был сгруппирован по столбцу "UUID студента" и для каждой колонки было получено количество уникальных значений. Назовём эту таблицу "nuniques". После этого была создана таблицу. В первом столбце - название колонки, в другом - количество уникальных значений в этой самой колонке из "nuniques", в третьем столбце перечислены все уникальные значения, количество которых посчитаны в предыдущем столбце. Эта таблица была использована для проверки наличия несостыковок в данных, например, указан разный пол у одного и того же студента в разных записях. Никаких подобных отклонений выявлено не было. Так эта таблица иногда использовалась для определения дальнейших действий со столбцами. Например, для выявления константных или бинарных признаков и т.д.

4. Было выявлено, что датасет не актуализирован, то есть в нём не обновлена информация за последний семестр (осень-зима 2024). Так же было принято сделать два датасета, один для абитуриентов и второй для студентов.

Работа с датасетом для предсказывания статуса студентов по баллам за предметы с 1 семестра

1. Была изменена таргетная переменная у тех студентов, которые были отчислены во время или после 3 семестра, так как отсутствовала полная информация об семестре.

2. Было принято решение использовать информацию о студентах, поступивших в 2023 году, хотя в выборке были студенты, поступившие и в 2022, и в 2024 годах. Это было сделано по двум причинам. Во-первых, взять информация о студентах, поступивших в 2024 году, не была актуализирована на конец 2024 - начало 2025 годов (неизвестно, кто из этих студентов отчислен, а кто активен). Во-вторых, из студентов 2022 года известны только активные, среди них нет ни одного отчисленного. Такие данные не имеют никакой ценности. Далее были удалены любые записи, не связанные с 1 семестром. Необходимо будет предсказывать, отчислится ли студент после 1 семестра (по результатам первого семестра, во время второго семестра или по результатам второго семестра). Так же были удалены записи о пересдачах предметов во время второго семестра, так как баллы определяются только по первому семестру.

3. Были удалены все записи о невыбранных предметах или о предметах, которые не были выполнены по уважительным причинам. Далее были заполнены пропуски в баллах. Так как пропуски в столбце "Оценка" отсутствуют, использовались его значения для определения баллов для вставки в соответствующий столбец "Баллы" ('зач.': 70, 'неуваж.': 0, 'недсд.': 0, '4': 70, 'недоп.': 0, '5': 90, 'незач.': 20, '2': 20, '3': 50). После заполнения пропусков в баллах, посчитано общее количество пересдач у каждого студента и была сохранена только одна запись с максимальными баллами за все сдачи. После этого добавлен признак, характеризующий общее количество долгов по итогам 1 семестра. Далее была применена формула к оставшимся записям о баллах, вычисляющая рейтинг студента. Эти значения были добавлены как новый признак.

4. После этого были выполнены те же действия, что и для датасета для абитуриентов, начиная с 3 пункта.

Работа с датасетом для предсказывания статуса абитуриентов по баллам вступительных испытаний

1. Было принято решение использовать информацию о студентах, поступивших в 2023 году, хотя в выборке были студенты, поступившие и в 2022, и в 2024 годах. Это было сделано по двум причинам. Во-первых, взять информация о студентах, поступивших в 2024 году, не была актуализирована на конец 2024 - начало 2025 годов (неизвестно, кто из этих студентов отчислен, а кто активен). Во-вторых, из студентов 2022 года известны только активные, среди них нет ни одного отчисленного. Такие данные не имеют никакой ценности. Далее были удалены любые записи, не связанные с 1 семестром. Необходимо будет предсказывать, отчислится ли студент после 1 семестра (по результатам первого семестра, во время второго семестра или по результатам второго семестра). Так же были удалены записи о пересдачах предметов во время второго семестра, так как баллы определяются только по первому семестру.

2. Была добавлена новая таргетная колонка. Туда скопированы все базовые значения таргетов и изменены таргетные переменные на «Активные» у тех студентов, у которых написано, что они отчислились в течении 3 семестра или по его итогам, так как данные о третьем семестре неполные.

3. Удаление лишних колонок.

"Наименование олимпиады" – удалена, так как слишком много различных значений, заместо этой колонки будет использована колонка "Тип олимпиады"

"Категория конкурса БВИ" - оставлено, но убраны значения, связанные со льготами, так как для этого будет использоваться отдельный столбец. В данном столбце же остались только олимпиады. Наличие любой олимпиады объединено в одно значение.

"UUID оценки" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Контрольное мероприятие" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Тип оценки" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Наименование дисциплины" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Модель реализации" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Оценка" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Балл" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Дата получения оценки" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Семестр" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Сквозной семестр" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Год" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Наименование дисциплины ИТС" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"ИТС наименование в дипломе" - удалено, так как не требуются колонки, которые хоть как-то связаны с предметами.

"Наименование напр. подготовки" - удалено, так как содержит слишком много различных значений, заместо этой колонки будет использована колонка "Код направления подготовки"

"Образовательная программа" - удалено, так как содержит слишком много различных значений, заместо этой колонки будет использована колонка "Код направления подготовки"

"ЛН студента" - удалено, так как эта колонка нам неинформативна

"ЛН студента (архив)" - удалено, так как эта колонка нам неинформативна

"Формирующее подразделение" - удалено, заместо этой колонки будет использована колонка "Уровень подготовки"

"Год начала обучения" - удалено, так как константный признак

"Срок освоения" - удалено, заместо этой колонки будет использована колонка "Уровень подготовки"

"Область образования" - удалено, так как константный признак

"Целевик" - удалено, заместо этой колонки будет использована колонка "Вид конкурса"

"Хоз. группа | Вкл. обучение" - удалено, так как константный признак

"Малоимущий" - удалено, так как константный признак

"Слушатель" - удалено, так как константный признак

"Курс" - удалено, так как курс не важен

"Состояние студента" - удалено, так как уже есть новая колонка с таргетом

"Активное состояние" - удалено, так как это повторение таргета

"В архиве" - удалено, так как это сильная подсказка для таргета (по сути аналог таргета)

"Приказ об отчислении" - удалено, так как это повторение таргета

"Акад. отп. по сост. здор." - удалено, так как это повторение таргета

"Акад. отп. по уходу за реб." - удалено, так как это повторение таргета

"UUID персоны" - удалено, так как эта колонка нам неинформативна

"Место рождения" - удалено, так как есть колонки, которые будет легче использовать для получения информации о национальности и т.д.

4. Теперь, после удаления лишних колонок необходимо сгруппировать записи по "UUID студента", так как остались только те признаки, которые не меняются. После группировки "UUID студента" была удалена.

5. Далее точечная обработка признаков. Первые и третьи числа в коде направления были выделены в отдельные признаки (2 число не нужно, так как оно означает уровень подготовки). "Код направления подготовки" была удалена. После этого был выполнен one-hot-encoding для колонки «Основа освоения (сокр.)» и она была удалена с одной из полученных колонок, чтобы избежать мультиколлинеарности. Такой же алгоритм был выполнен для следующих колонок: «Нуждается в общежитии», "Иностранный абитуриент (МОН)", "Пол". У каждого студента было определено количество полных лет на момент поступления. Страны из признака "Гражданство" были разделены на Россию, пост-советское пространство и другие. Для полученных категорий так же был выполнен one-hot-encoding. Добавлен признак, объясняющий, является ли городом регистрации студента Екатеринбург. Добавлен признак, объясняющий, является ли областью регистрации студента Свердловская область. Объединены различные законченные учебные заведения по типам ((не указано), Вечерняя школа, Основная общеобразовательная школа, Средняя общеобразовательная школа) -> Школа, (Гимназия, Лицей, Дневная школа, лицей, гимназия) -> Профильная Школа, (Кадетский корпус, Суворовское училище) -> Военное уч. заведение, (Колледж, Техникум, Профессиональное училище) -> СПО. Так же был выделен признак, показывающий, сколько лет прошло с момента окончания учебного заведения до поступления в университет. Объединены льготы по категориям (Сироты, Инвалиды, Боевые действия, Квота для иностранных граждан). В качестве нового признака добавлен индекс человеческого развития в стране, из которой прибыл студент. Был проведён one-hot-encoding по всем оставшимся и появившимся категориальным колонкам. Заменены значения таргета (Отп.акад., Отп.б.посещ., Отчислен) -> 1, (Активный, Зак. с дипломом) -> 0.

6. Полученный датасет был разделён на два. В одном содержатся только магистры (671 запись, 593/78), в другом только бакалавры и специалисты (1612 записей, 1392/220).