Федеральное государственное автономное

образовательное учреждение

высшего образования

«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт космических и информационных технологий

институт

Кафедра информатики

кафедра

**ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ №2**

Инженерия машинного обучения

тема

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Преподаватель |  |  |  |  | Е.О. Пересунько |
|  |  |  | подпись, дата |  | инициалы, фамилия |
| Студент | КИ19-16/2Б, 031944771 |  |  |  | К.Е. Бояршинов |
|  | номер группы, зачетной книжки |  | подпись, дата |  | инициалы, фамилия |

Красноярск 2022

# СОДЕРЖАНИЕ

[1 Цель 3](#_Toc116150574)

[2 Задание 3](#_Toc116150575)

[3 Описание решения 3](#_Toc116150576)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 6](#_Toc116150577)

## Цель

Знакомство с инструментами организации процесса MLOps, получение практических навыков по работе с версионированием экспериментов и автоматизации пайплайнов машинного обучения.

## Задание

1. Создать удаленный репозиторий для проекта.

2. Разработать решение задачи в соответствии с вариантом.

a. Построить автоматический ML-пайплайн, состоящий из всех необходимых шагов для решения задачи (предобработка данных, обучение моделей, оценка качества и т.д.).

b. Провести серию экспериментов, используя как минимум три различные модели машинного обучения для решения поставленной задачи. При защите работы необходимо продемонстрировать возможность перемещения по истории экспериментов с целью выбора конкретной модели.

3. Выполнить аугментацию, сгенерировав дополнительные данные с помощью различных преобразований, и сохранить новый датасет.

4. Повторить пункт 2, используя все данные (исходные и аугментированные).

5. В отчете привести сравнение результатов, получаемых при различных условиях экспериментов на разных наборах данных (с аугментацией и без).

## Описание решения

В рамках работы задачей было выбрано прогнозирование неуспеваемости в прошлых учебных годах среди студентов. Данные взяты с платформы Kaggle (Рисунок 1):

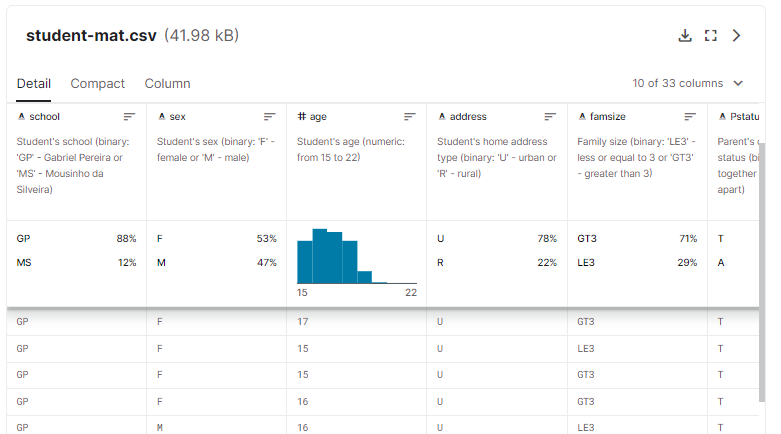


Рисунок 1 – Формат датасета

Для решения задачи регрессии были выбраны Логистическая Регрессия, Случайный лес и Стохастический градиентный спуск.

Решение задачи разбивается на 4 этапа:

- загрузка датасета,

- выделение признаков,

- обучение модели,

- оценка результатов.

Каждый из этапов является отдельным скриптом и частью выстроенного с помощью dvc пайплайна (Рисунок 2).

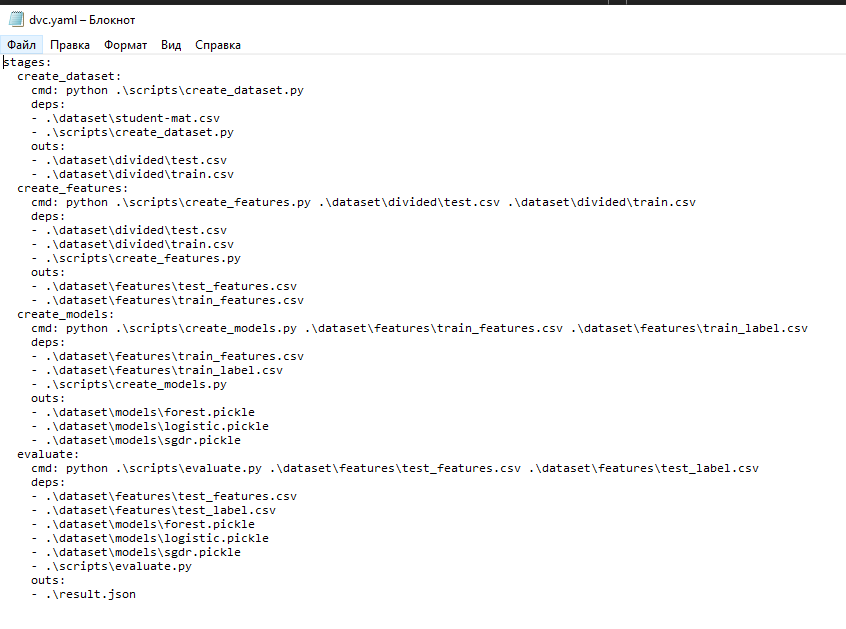


Рисунок 2 – Автоматический ML-пайплайн (файл dvc.yaml)

Для обучения модели были выбраны следующие признаки: ["age", "Medu", "Fedu"].

В рамках аугментации в датасет и набор признаков для обучения был добавлен столбец Mid\_Edu, рассчитанный как cреднее значение между образованием матери и образованием отца.

Для оценки работы алгоритмов были выбраны метрики R-squared и RMSE.

Ниже представлен вывод метрик для запуска трех алгоритмов без аугментации и с аугментацией (Рисунок 3).

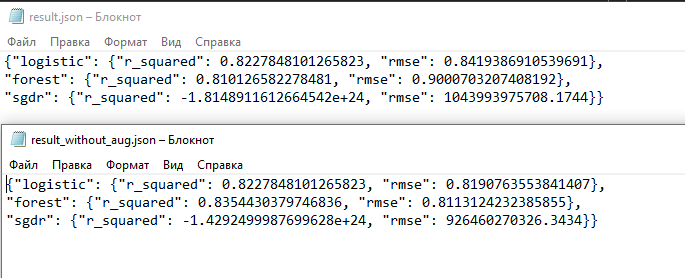


Рисунок 3 – Полученные метрики

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. СТУ 7.5–07–2021 Система менеджмента качества. Общие требования к построению, изложению и оформлению документов учебной деятельности. – Введ. 7.12.2021. – Красноярск: ИПКСФУ, 2021.-61 с.