

# Прогнозирование статуса студента

Цифровой прорыв 2022

# Суть задачи

- Определить, что стало со студентом:
  - отчислился
  - ушёл в академ. отпуск
  - успешно выпустился
- Данные:
  - 13584 строки в тренировочном датасете
  - 24 поля, включая ID






# Описание выбранной модели

- CatBoost ([arXiv:1706.09516](https://arxiv.org/abs/1706.09516))

градиентный бустинг - часто самое эффективное решение (тем более для таких табличных данных)

catboost хорошо работает с категориальными полями

низкий порог входа - даже default версия “из коробки” работает хорошо

	CatBoost	LightGBM	XGBoost	H2O
	Default	Default	Default	Default
 Adult	0.27298	0.28716 +5.20%	0.28009 +2.61%	0.27607 +1.14%
 Amazon	0.13811	0.16716 +21.04%	0.16536 +19.74%	0.16950 +22.73%
 Click prediction	0.39112	0.39749 +1.63%	0.39764 +1.67%	0.39785 +1.73%
 KDD appetency	0.07138	0.07482 +4.82%	0.07466 +4.60%	0.07355 +3.05%
 KDD churn	0.23193	0.23565 +1.61%	0.23369 +0.76%	0.23287 +0.41%

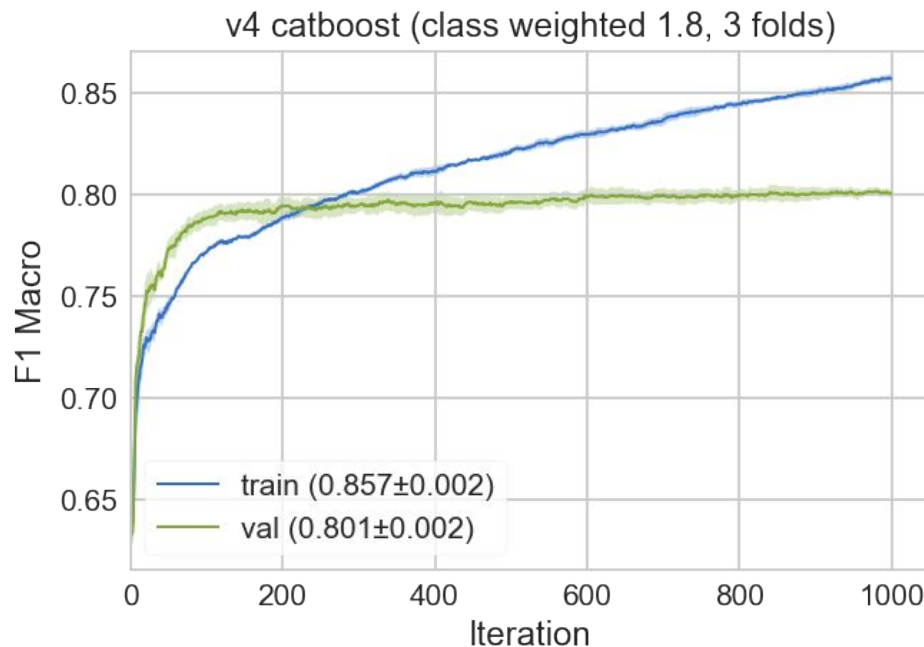
Benchmark с официальной страницы [catboost.ai](https://catboost.ai).  
Метрика LogLoss (меньше - лучше)

# Точность решения

- используем метрику F1-macro

$$F1_{macro} = \frac{\sum_1^N F1_i}{N}$$

классы влияют одинаково на метрику, но в датасете они не сбалансированы →  
файнтюнинг модели  
взвешиванием классов



Результаты финальной модели в процессе обучения на кросс-валидации (3 folds).  
Итоговый результат в скобках в легенде

# Выводы по данным

- данные грязные, необходима предобработка
- множество полей мало влияющих на результат:
  - Общежитие, Наличие родителей, Изучаемый язык, Пол, Село, Опекунство, Пособие и т.п.
- интуитивно значимыми полями могут быть:
  - Средний доход семьи, Приоритет направления при поступлении

```
Код_группы      20606
Год_Поступления 2016
Пол             Муж
Основания       БН
Изучаемый_Язык  Английский язык
Дата_Рождения   1997-11-14 00:00:00.000
Уч_Заведение    МБОУ Лицей №101
Где_Находится_УЗ  Россия, Алтайский край, г Барнаул
Год_Окончания_УЗ 2016.0
Пособие         0.0
Страна_ПП       Россия
Регион_ПП       Алтайский край
Город_ПП        Барнаул
Общежитие       0.0
Наличие_Матери  1
Наличие_Отца    1.0
Страна_Родители Россия
Опекунство      0.0
Село            0.0
Иностранец      0.0
КодФакультета   25.0
СрБаллАттестата 77.0
Name: 75771, dtype: object
```

```
Код_группы      20606
Год_Поступления 2016
Основания       бн
КодФакультета   25
СрБаллАттестата 77.0
Код1            20
Код2            60
Код3            606
Муж             1
Год_Рождения    1997
Возраст_Поступления 19
Перерыв         0
ПолнаяСемья    True
Учеба           ш
МестоЖит        брн
Name: 75771, dtype: object
```

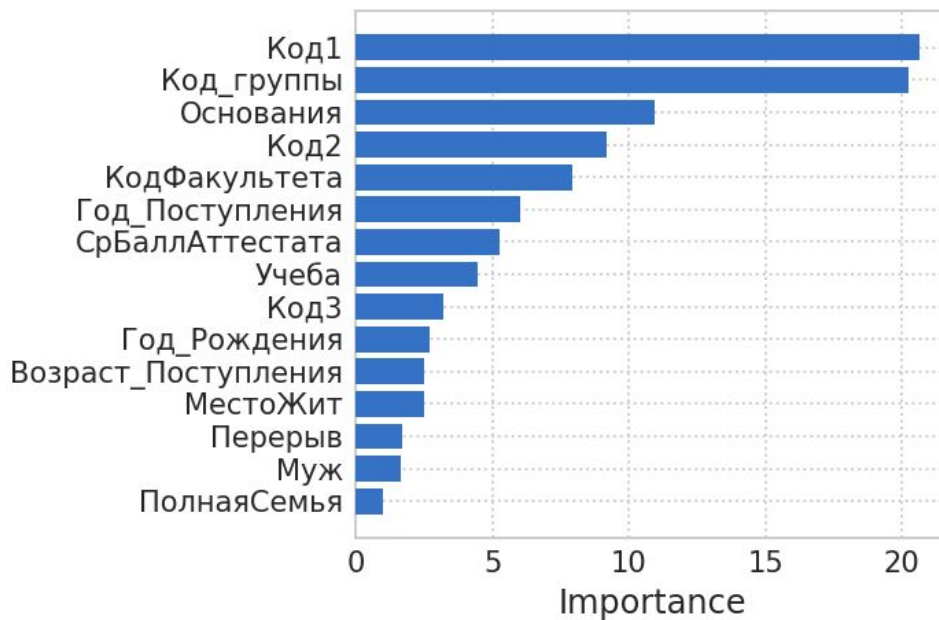
Одна и та же строка данных до и после обработки

# Выводы по данным

- Наибольшая важность - код группы

получается, что вероятность успешного выпуска по больше зависит от группы, куда попадает студент (видимо это связано с преподавателями в них), чем от балла аттестата или школы

хорошо было бы понимать, как кодифицируется группа (обычно в номере группы больше информации чем кажется).



Важность фичей.

Код1 - 1 и 2 цифры кода группы,  
Код2 - 3 и 4; Код3 - 3, 4 и 5 цифры

# Контактные данные

e-mail: [nikita.petrov.1997@gmail.com](mailto:nikita.petrov.1997@gmail.com)