



## Задача «Прогнозирование статуса студента»

### Введение:

По данным Минобрнауки РФ в этом году более 6,5 млн. человек подали заявление на зачисление в учебное заведение. В 2022 г. количество бюджетных мест увеличилось до 588 044, большинство из них — в региональных университетах.

Высшее образование имеет большое значение для формирования карьерной траектории. На сегодняшний день в России около 700 вузов предоставляют возможность обучения по образовательным программам в различных направлениях. После школы или колледжа молодые люди выбирают, в какой сфере хотят развиваться. И, конечно, каждый будущий студент хочет получить действительно качественное и разностороннее образование.

Ежегодно тысячи абитуриентов в нашей стране участвуют в приемной кампании, выбирая себе будущую профессию или дополнительные знания и навыки, чтобы повысить свои профессиональные компетенции и стать востребованным специалистом. Но все ли студенты заканчивают университет?

Перед участниками чемпионата стоит задача предсказать будущий потенциальный статус студента на основе данных нескольких тысяч студентов. Такая прогнозная модель может стать хорошим инструментом при планировании учебной работы в вузах, взаимодействии с работодателями или реализации научных проектов.

### Условие задачи:

На основе данных об абитуриенте участникам чемпионата необходимо разработать модель, которая будет предсказывать текущий статус студента, а именно:

- продолжит ли студент обучение
- отчислится
- возьмет академический отпуск

### Описание входных значений:

**train.csv** — файл, содержащий данные студентов для тренировки модели машинного обучения;

**test.csv** — файл с данными для предсказания;

**submission.csv** — пример файла для отправки.

### Описание полей данных:

ID: персональный идентификатор абитуриента

Код\_группы: идентификатор группы

Год\_Поступления: год поступления абитуриента

Пол: пол абитуриента, имеет идентификаторы:

- Муж
- Жен

Основания: основания поступления абитуриента. Имеются следующие значения:

- СН - специальный набор
- ЦН - целевой набор
- БН - бюджетный набор
- ОО - общие основания
- ДН - дополнительный прием

Изучаемый\_Язык: иностранный язык, который изучался абитуриентом. Поскольку долгое время данное поле было полем ввода, то происходит дублирование значений. (первая фишка)

Дата\_Рождения: дата рождения абитуриента

Уч\_Заведение: наименование учебного заведения

Где\_Находится\_УЗ: локация учебного заведения. Структура данного поля разнообразна, может происходить наименование города, а может указываться область+город или даже страна+область+город. (вторая фишка)

Год\_Окончания\_УЗ: год окончания учебного заведения

Пособие: получает ли абитуриент пособия. Возможные значения:

- 0 - абитуриент не получает пособие
- 1 - абитуриент получает пособие

Страна\_ПП: страна постоянного проживания. Поскольку долгое время данное поле было полем ввода, то происходит дублирование значений. (третья фишка)

Регион\_ПП: регион постоянного проживания. Поскольку долгое время данное поле было полем ввода, то происходит дублирование значений. (четвертая фишка)

Город\_ПП: город постоянного проживания. Поскольку долгое время данное поле было полем ввода, то происходит дублирование значений. (пятая фишка)

Общежитие: требуется ли общежитие абитуриенту. Возможные значения:

- 0 - абитуриенту не требуется общежитие
- 1 - абитуриенту требуется общежитие

Наличие\_Матери: поле, показывающие наличие матери. Возможные значения:

- 0 - у абитуриента нет матери
- 1 - у абитуриента есть мать

Наличие\_Отца: поле, показывающие наличие отца. Возможные значения:

- 0 - у абитуриента нет отца
- 1 - у абитуриента есть отец

Страна\_Родители: страна постоянного проживания родителей

Опекунство: является ли абитуриент опекаемым. Возможные значения:

- 0 - абитуриент не является опекаемым
- 1 - абитуриент является опекаемым

Село: проживает ли абитуриент в сельской местности. Возможные значения:

- 0 - абитуриент не проживает в сельской местности
- 1 - абитуриент проживает в сельской местности

Иностранец: является ли абитуриент иностранным студентом. Возможные значения:

- 0 - абитуриент не является иностранцем

- 1 - абитуриент является иностранцем

КодФакультета: идентификатор факультета

СрБаллАттестата: средний балл аттестата или средний балл ЕГЭ. Зависит от того, сдавал ли абитуриент ЕГЭ

Статус (целевая переменная): статус студента. Возможные значения:

- -1 – ушел в академический отпуск
- 4 – закончил обучение
- 3 – отчислен
- 5 – призван в армию

### Метрика:

В качестве метрики выступает F1-мера столбцу Статус:

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

**Recall** считается как:

$$Recall = \frac{TP}{TP + FN}$$

где: TP (True Positive) - количество верно угаданных значений одного класса  
FN (False Negative) - количество не правильно угаданных значений одного класса

**Precision** считается как:

$$Precision = \frac{TP}{TP + FP}$$

где: TP (True Positive) - количество верно угаданных значений одного класса  
FP (False Positive) - количество всех неправильно классифицированных значений одного класса

### Правила чемпионата:

1. С момента открытия датасета до момента завершения приема решений репозиторий участника, в котором он ведет разработку по задаче текущего чемпионата, должен оставаться закрытым.

2. Участник обязан открыть доступ к репозиторию на чтение по ссылке (которая была прикреплена в ЛК в поле «Ссылка на код (гитхаб)») не позднее чем в течение 12 часов с момента окончания дедлайна отправки решений на региональном чемпионате.

3. Согласно п. 5.8 Положения в процессе верификации решений организаторы и технические эксперты, проверяющие решения участников, вправе назначить интервью с участниками чемпионата. Участник получит приглашение и ссылку на интервью не позднее чем за 12 часов до публикации итогового лидерборда. Пропуск интервью участником является поводом для дисквалификации.

4. Организаторы вправе исключить участника из призовых позиций лидерборда за непредоставление одного из артефактов решения задачи: тизера, скринкаста, презентации, ссылки на репозиторий.

5. Организаторы вправе дисквалифицировать участника в случае выявления плагиата кода или несоблюдения Положения конкурса.

6. Участник, получивший 2 дисквалификации за сезон проекта, попадает в чёрный список с дальнейшим отстранением от участия в чемпионатах до конца сезона.