

Искусственный интеллект - тимлид

В компании, специализирующейся на разработке программного обеспечения, внедрена система планирования времени сотрудников. Команды работают по двухнедельным спринтам и записывают время, потраченное на те или иные задачи. Необходимо отметить также, что записываются все рабочие задачи, включая встречи и технические собеседования.

Все данные по всем проектам компания собирает и хранит. Включая стендапы команд. Суть предлагаемой вам задачи заключается в оценке времени, которое понадобится на разработку того или иного функционала в проекте. Программа должна работать как технический лидер проекта, распределяющий и оценивающий время на выполнение задач для своих сотрудников. Для этого вам предлагаются исторические данные, на основе которых и предлагается сделать оценку.

Условие задачи

На основе личных параметров тимлидов, ответственных разработчиков, описания задачи в спринте и комментариев к ней - разработайте модель которая, сможет оценить время которое уйдет на задачу.

Описание входных значений

- train/train_issues.csv — содержит в себе 9589 различных задач в спринте;
- train/train_comments.csv — содержит комментарии разработчиков к задачам обучающего набора;
- test/test_issues.csv — содержит в себе 1070 различных задач для которых требуется предсказать потрачено команд;
- train/test_comments.csv — содержит комментарии разработчиков к задачам тестового набора;
- sample_solution.csv — пример файла для отправки;
- employees.csv - список работников и их контактная информация.

Дадим пояснение некоторым столбцам в данных:

1) issues.csv

- id - идентификатор задачи в глобальной базе данных
- created - дата создания задачи
- key - ключ задачи, используется для идентификации задачи внутри проекта
- summary - описание задачи

- project_id - идентификатор проекта, по которому выполняется задача
- assignee_id - идентификатор сотрудника, на которого назначена задача (таблица сотрудников)
- creator_id - идентификатор сотрудника, который создал задачу (таблица сотрудников)
- overall_worklogs - количество времени в секундах, ушедшего на решение задачи

2) issues.csv

- comment_id - идентификатор комментария
- text - текст комментария
- issue_id - идентификатор задачи (таблица задач)
- author_id - идентификатор автора комментария (таблица сотрудников)

Метрика

В качестве метрики выступает R^2 .

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

SS_{res} - сумма квадратов остаточных ошибок.

SS_{tot} - общая сумма ошибок.