цифровой прорыв
сезон: ИИ

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

Задача «Разработка алгоритма прогнозирования выполнения задачи»

Введение

В процессе разработки программного обеспечения часто можно столкнуться с проблемами качества, стоимости и надёжности. Все эти компоненты оцениваются в первую очередь в трудозатратах. Некоторые программы содержат миллионы строк исходного кода, которые, как ожидается, должны правильно исполняться в изменяющихся условиях. Разработка программных решений — это процесс, сравнимый по сложности с созданием и сборкой самолёта. И, как и любой процесс, даже самый сложный, разработку можно разделить на части.

Для организации непрерывного процесса разработки сложных программных продуктов требуется тщательное планирование работы. Основной инструмент — это декомпозиция разработки программного продукта на мелкие шаги с последующей оценкой качества выполненной работы и временных затрат на исполнение.

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 предоставляются 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 будет затрачено на выполнение задачи.

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;
train/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 сотрудники)

Метрика

В качестве метрики выступает R2.

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

SS res - сумма квадратов остаточных ошибок.
SS tot - общая сумма ошибок.

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 выявление плагиата кода или несоблюдение Положения конкурса.

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