Контрольное задание

**Общее описание**

В этом задании мы с вами познакомимся с таким понятием, как контрфактическое моделирование. Часто этот термин используют историки, когда пытаются ответить на вопрос «могли ли события пойти иначе?».

**Проблематика.** Оценка повышения энергоэффективности зданий может быть сложной задачей, поскольку невозможно точно определить, сколько энергии здание использовало бы до капитального ремонта, в рамках которого было повышение энергоэффективности. Проведение A/B тестирования (т.е. постройка нескольких одинаковых зданий и капитальный ремонт части их них) экономически не может быть обосновано.

**Предполагаемый подход.** Одним из лучших решений в этой области, которые мы можем сделать, — это построить контрфактические модели. После капитального ремонта здания у нас появляется возможность сравнивать новое (более низкое) энергопотребление со смоделированными значениями для исходного здания. Это позволит нам рассчитать экономию от оптимизации. Более точные модели могли бы поддерживать лучшие рыночные стимулы и обеспечить более дешевое финансирование.

**Способ решения.** В этом задании мы будем стараться строить эти гипотетические модели для четырех типов энергии (электричество, холодная и горячая вода и пар) на основе исторических показателей использования этих видов энергии и наблюдаемой погоды.

**Важное замечание.** Не все здания могут обладать всеми типами счетчиков, часть данных может теряться, какие-то счетчики могли появиться в определенных зданиях уже во время наблюдений.

Контрольное задание

**Модуль 1. Загрузка, предобработка и подготовка данных**

В данном модуле участники должны загрузить данные, оптимизировать занимаемый объем данных, выполнить минимальную предобработку и подготовить отчет о проделанной работе.

Входные данные:

1. train.csv, test.csv, wtrain.csv, wtest.csv, objects.csv.

Подробное описание задания:

1. Загрузить данные (должны быть загружены все представленные данные, при необходимости исключена техническая информация, код должен содержать обработку исключений).
2. Произвести оптимизацию объемов занимаемой памяти без потери данных (проверены все типы столбцов всех загруженных данных, произведена оптимизация объемов занимаемой памяти данными, отображена информация об объемах памяти до оптимизации, после и процент оптимизации).
3. Подготовка датасета (при необходимости произведено объединение таблиц и аргументировано это объединение или аргументировано отсутствие необходимости выполнять эту работу на данном этапе, полученные данные сохранены на диск).
4. Подготовка отчет по модулю (отчет должен быть в формате HTML/PDF на выбор, отчет должен быть структурирован, описывать ход работы, содержать комментарии).

Выходные данные по модулю:

1. Отчет в формате HTML/PDF.
2. Файл .ipynb, полностью исполняемый и без ошибок.
3. Один или несколько итоговый файлов, представляющих оптимизированные и объединенные (при необходимости) данные.

Формат именования файлов:

M1\_{lastname}.{fileformat}

(например, M1\_Ivanov.html, M1\_Ivanov.ipynb)

**Приложение А. Описание структуры данных**

**train.csv**

|  |  |
| --- | --- |
| obj\_id | Ключ здания |
| counter | Тип счетчика (0 – электричество, 1 – холодная вода, 2 – горячий пар, 3 – горячая вода) |
| datetime | Метка времени снятия показаний счетчиков |
| value | Целевая переменная |

**building\_meta.csv**

|  |  |
| --- | --- |
| city | Ключ города/локации |
| obj\_id | Ключ здания |
| function | Назначение здания |
| square | Общая площадь здания |
| year | Год постройки здания |
| floors | Количество этажей в здании |

**weather\_[train/test].csv**

|  |  |
| --- | --- |
| city | Ключ города/локации |
| datetime | Метка времени измерения |
| temp | Температура воздуха |
| clouds | Облочность |
| dew | Точка росы |
| precipitation | Количество осадков |
| pressure | Давление на уровне моря |
| wdirection | Направление ветра |
| wspeed | Скорость ветра |

**test.csv**

|  |  |
| --- | --- |
| obj\_id | Ключ здания |
| counter | Тип счетчика (0 – электричество, 1 – холодная вода, 2 – горячий пар, 3 – горячая вода) |
| datetime | Метка времени снятия показаний счетчиков |

**predict.csv** – результирующий файл от участников

|  |  |
| --- | --- |
| value | Предсказанное значение |

ВНИМАНИЕ: последовательность в выходном файле должна быть строго соблюдена с тем, что было в test.csv.