Направление Data Scientist, Schlumberger

Аннотация
Добро пожаловать на виртуальную стажировку **Shift + Enter** от Schlumberger. Решая задания, ты узнаешь, из чего состоит работа начинающего специалиста по данным.
В твои обязанности как Junior Data Scientist входит:

- Реализация полного пайплайна Data Science : приготовление данных, обучение модели и оценка ее качества.
- Знание основных моделей машинного обучения и умение применять их на практике.

Развиваемые компетенции
По результатам выполнения заданий ты научишься:

1. Заполнять пропущенные значения в данных.
2. Делать data-driven[1]-выводы.
3. Строить предсказательную модель и оценивать ее точность.

А также сможешь прокачать следующие навыки: критическое мышление, машинное обучение и анализ данных.

Описание подзадач
Тебе предстоит изучить данные, содержащие параметры резервуара и работ по гидроразрыву пласта (ГРП)[2]. Технология основана на закачивании в пласт жидкости гидроразрыва, содержащей проппант[3], при давлении, достаточном для создания трещин гидроразрыва.

Предложи подход для заполнения пропущенных данных, проверь гипотезу о связи некоторых переменных и предскажи массу проппанта, которую следует закачать в рассматриваемый пласт.

Выполнение всего блока заданий займет у тебя не более 80–100 минут.

Рекомендуемый тайминг

1. 20–25 минут на первое задание.
2. 20–25 минут на второе задание.
3. 40–50 минут на третье задание.

Информация о загрузке решения
Этот проект содержит несколько подзадач. Можно загрузить файл, содержащий решение части заданий, но по возможности постарайся сделать их все.

Желаем удачи!

Задание 1. Заполни все недостающие данные

[1] Data-driven-подход ставит во главу угла анализ данных.
[2] Гидроразрыв пласта — широко применяемая технология для увеличения нефтеотдачи пласта.
[3] Проппант — гранулообразный материал, препятствующий закрытию трещин после окончания работы по ГРП и создающий высокопроводящий канал для притока нефти из пласта в скважину.

Утром ты обнаружил на почте письмо от руководителя IT-команды Schlumberger.

Привет!

У нас есть таблица (ссылка на скачивание – во вложении 1), содержащая некоторые параметры резервуара и проведенных работ по ГРП. Мы собирали данные из разных частей света, от разных команд, поэтому часть данных не была записана либо потерялась. Мы не можем позволить себе исключать наблюдения из-за пропусков. Поэтому заполни, пожалуйста, все пропущенные значения. Попробуй разные способы и выбери наиболее подходящий.

Оформи свое решение и пришли его в течение двух часов. Спасибо!

--
С уважением,
Михаил Петров[4]

Полезные материалы

- Статья о том, как работать с пропущенными значениями в Python.

Формат конечного результата

Файл формата .ipynb/.py.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.

[4] Все имена и названия вымышленные, любые совпадения случайны. Данные задания могут быть изменены в целях конфиденциальности.

Вложение 1. Дата-сет

Скачать данные можно, перейдя по ссылке: dataset.

Описание дата-сета

- Well_id – уникальный идентификатор скважины.
- Reservoir pressure – давление жидкости в пласте, [давление].
- ISIP (Instantaneous Shut-In Pressure) – давление жидкости после мини-ГРП[5], [давление].
- Closure Pressure – давление жидкости, когда трещина полностью закрывается, [давление].
- PAD Volume – объем чистой жидкости, закачанной в трещину, [объем].
- Fluid Efficiency – эффективность жидкости гидроразрыва (вычисляется как отношение объема трещины к объему закачанной жидкости), [0-1].
- Transmissibility (пропускная способность) – коэффициент, характеризующий сопротивление жидкости течению через пласт.
- Total Prop Mass – общая масса закачанного проппанта во время основной обработки, [масса].
- Max Prop Conc – максимальная концентрация проппанта во время основной обработки, [концентрация].

[5] Небольшой гидроразрыв пласта, выполняемый перед основным ГРП, для получения критически важных данных о проектировании и выполнении работ и подтверждения прогнозируемого отклика интервала обработки.

Задание 2. Проверь гипотезу о связи данных

Твой руководитель остался доволен результатами проделанной работы и направил тебе письмо с новой задачей, связанной со знанием математической статистики.

Привет!

Спасибо, что заполнил пропущенные данные.

Наш коллега, занимающийся созданием планов проведения работ по гидроразрыву пласта, однажды поделился своим эмпирическим правилом (rule of thumb): чем больше давление в резервуаре[6], тем больше надо закачать в трещину проппанта.

Получится ли проверить его утверждение, используя имеющийся дата-сет? Сможешь ли ты построить собственное эмпирическое правило, основанное на данных?

Hints: Найди, какие характеристики дата-сета сильно взаимосвязаны между собой, и предложи свое правило исходя из этого.

Жду файл с кодом, содержащий пояснения по правилу коллеги, и твои рекомендации по данным (можешь дополнить код из предыдущей задачи).

Спасибо!

--
С уважением,
Михаил Петров

Полезные материалы

- Статья, описывающая построение тепловой карты корреляции для нахождения взаимосвязи между параметрами.

Формат конечного результата

Файл в формате .ipynb/.py, содержащий выводы и рекомендации.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.

[6] (Природный) резервуар углеводородов — состоящее из коллектора тело горных пород, частично или со всех сторон ограниченное относительно непроницаемыми границами, выступающее как естественное вместилище для нефти, газа и воды.

Задание 3. Натренируй модель машинного обучения

Твоим финальным заданием станет оценка точности модели, которую тебе нужно построить для предсказания количества проппанта, нужного для проведения ГРП.

Перед тем как уйти на еженедельный синхрон с командой, ты увидел пояснения к новой задаче в почте.

Привет!

Нам снова нужна твоя помощь: необходимо примерно оценить объем предстоящей работы по ГРП. Времени на полноценные расчеты и оптимизацию нет, поэтому мы бы хотели использовать силу машинного обучения.

Опираясь на имеющиеся данные, построй модель для предсказания массы проппанта Total Prop Mass и оцени ее точность.

Hints: Для тренировки модели предсказания массы проппанта используй любые доступные колонки из дата-сета.

Жду файл с кодом, большое спасибо за подключение к проекту!

--
С уважением,
Михаил Петров

Полезные материалы

- Статья, содержащая информацию о популярных алгоритмах машинного обучения.
- Инструкция о том, как строить ML-прогнозы.

Формат конечного результата

Файл в формате .ipynb/.py, содержащий код и пояснения к нему.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.