Описание проекта

Допустим, вы работаете в добывающей компании «ГлавРосГосНефть». Нужно решить, где бурить новую скважину.
Шаги для выбора локации обычно такие:

- В избранном регионе собирают характеристики для скважин: качество нефти и объём её запасов;
- Строят модель для предсказания объёма запасов в новых скважинах;
- Выбирают скважины с самыми высокими оценками значений;
- Определяют регион с максимальной суммарной прибылью отобранных скважин.

Вам предоставлены пробы нефти в трёх регионах. Характеристики для каждой скважины в регионе уже известны. Постройте модель для определения региона, где добыча принесёт наибольшую прибыль. Проанализируйте возможную прибыль и риски техникой *Bootstrap*.

Инструкция по выполнению проекта

1. Загрузите и подготовьте данные. Поясните порядок действий.
2. Обучите и проверьте модель для каждого региона:
   2.1. Разбейте данные на обучающую и валидационную выборки в соотношении 75:25.
   2.2. Обучите модель и сделайте предсказания на валидационной выборке.
   2.3. Сохраните предсказания и правильные ответы на валидационной выборке.
   2.4. Напечатайте на экране средний запас предсказанного сырья и *RMSE* модели.
   2.5. Проанализируйте результаты.
3. Подготовьтесь к расчёту прибыли:
   3.1. Все ключевые значения для расчётов сохраните в отдельных переменных.
   3.2. Рассчитайте достаточный объём сырья для безубыточной разработки новой скважины. Сравните полученный объём сырья со средним запасом в каждом регионе.
   3.3. Напишите выводы по этапу подготовки расчёта прибыли.
4. Напишите функцию для расчёта прибыли по выбранным скважинам и предсказаниям модели:
   4.1. Выберите скважины с максимальными значениями предсказаний.
   4.2. Просуммируйте целевое значение объёма сырья, соответствующее этим предсказаниям.
   4.3. Рассчитайте прибыль для полученного объёма сырья.
5. Посчитайте риски и прибыль для каждого региона:
   5.1. Примените технику *Bootstrap* с 1000 выборок, чтобы найти распределение прибыли.
   5.2. Найдите среднюю прибыль, 95%-й доверительный интервал и риск убытков. Убыток — это отрицательная прибыль.
   5.3. Напишите выводы: предложите регион для разработки скважин и обоснуйте выбор.

Описание данных

Данные геологоразведки трёх регионов находятся в файлах:

- `/datasets/geo_data_0.csv`. Скачать датасет
- `/datasets/geo_data_1.csv`. Скачать датасет
- `/datasets/geo_data_2.csv`. Скачать датасет
- *id* — уникальный идентификатор скважины;
- *f0, f1, f2* — три признака точек (неважно, что они означают, но сами признаки значимы);
- *product* — объём запасов в скважине (тыс. баррелей).

Условия задачи:

- Для обучения модели подходит только линейная регрессия (остальные — недостаточно предсказуемые).
- При разведке региона исследуют 500 точек, из которых с помощью машинного обучения выбирают 200 лучших для разработки.
- Бюджет на разработку скважин в регионе — 10 млрд рублей.
- При нынешних ценах один баррель сырья приносит 450 рублей дохода. Доход с каждой единицы продукта составляет 450 тыс. рублей, поскольку объём указан в тысячах баррелей.
- После оценки рисков нужно оставить лишь те регионы, в которых вероятность убытков меньше 2.5%. Среди них выбирают регион с наибольшей средней прибылью.

Данные синтетические: детали контрактов и характеристики месторождений не разглашаются.

Как будут проверять мой проект?

Мы подготовили критерии оценки проекта, которыми руководствуются ревьюеры. Прежде чем приступить к решению кейса, внимательно их изучите.
На что обращают внимание ревьюеры, проверяя проект:

- Как вы готовите данные к обучению?
- Выполнили все шаги по инструкции?
- Все ли условия бизнеса учтены?
- Какие выводы об исследовании задачи делаете?
- Корректно ли выполнена процедура *Bootstrap*?
- Предложен ли регион для разработки скважин? Обоснован ли выбор?
- Не дублируете ли код?
- Следите ли за структурой проекта и поддерживаете ли аккуратность кода?

Всё, что вам нужно знать, есть в шпаргалках и конспектах прошлых тем.
Успеха!