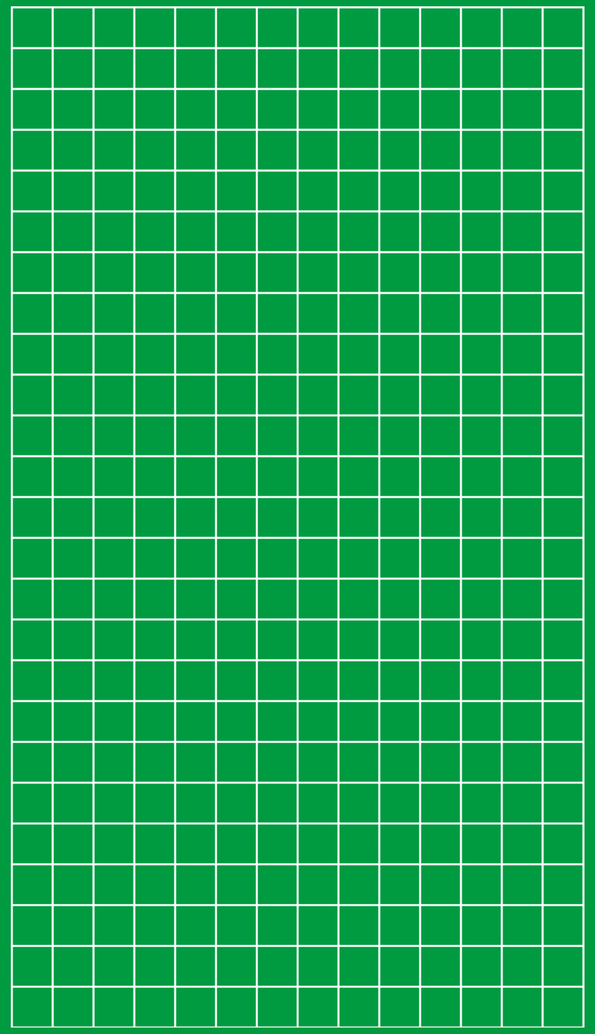
ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ



Введение в статистику

Проект #5

Ковчег

Кейс

- Компания планирует проводить тестовые запуски рекламы раз в день
- Нельзя, чтобы каждый клиент получал каждый день новое рекламное сообщение очередной рекламной кампании
- Тестовые запуски будут проводиться на маленьких выборках из всех клиентов
- Всего клиентов 80 тысяч, выборка размера 400
- Выборка должна быть репрезентативной, чтобы замеры на ней позволяли делать хотя бы какие-то выводы обо всех клиентах
- Проверку репрезентативности можно осуществить на основе одной кампании, которую проводили на всех 80 тысячах клиентов

Проблема

- Если один раз составить репрезентативную выборку и использовать ее во всех тестах, будем наблюдать эффект очередной рекламы на человека, которому каждый день показывают рекламу

Идея решения

- Разработать алгоритм, который будет формировать каждый раз новую выборку для тестового запуска очередной рекламной кампании

Ваша роль

- Аналитик отдела работы с клиентами

Ваша задача

- "Написать функцию на языке Python, которая формирует выборки
- Проверить, что функция формирует выборки репрезентативно
- Подготовить слайды, на которых объяснить способ формирования выборок и показать, насколько репрезентативные выборки он формирует"

Доступные данные

- Датасет ark.csv.
- Строки — клиенты
- Столбцы:
  - Id — идентификатор клиента,
  - Age — возраст,
  - Gender — пол,
  - Purhased — купил ли клиент товар после рекламной кампании, проводившейся для всех клиентов.

Ожидаемый результат

- Слайды в PowerPoint с презентацией результатов
- Jupyter Notebook с расчетами для слайдов

Требования к слайдам

Если слайды или Jupyter Notebook не приложен, решение кейса оценивается в 0 баллов

Понятность/внешний вид

- Внешний вид презентации не мешает воспринимать информацию
- Понятно на какие вопросы отвечает каждый слайд
- Содержимое таблиц, графиков понятно из слайда без необходимости открывать исходный датасет
- Выводы явно сформулированы

Максимум: 4 балла

Обоснованность

- Выводы основаны на таблицах, графиках, показателях, полученных из данных
- Таблицы и графики получены скриншотом или картинкой из Jupyter Notebook, поэтому их можно перепроверить
- Расчеты корректны

Максимум: 4 балла

Реакция заказчика (руководителя)

0 баллов:

Не принимает, ищет другого исполнителя

1 балл:

Частично принимает, считает необходимым отдать на доработку текущему исполнителю

2 балла:

Принимает, готов пересылать слайды от своего имени, под свою ответственность

Максимум: 2 балла

Синий уровень: подход к решению

Опишите на слайде способ получения простой случайной выборки

На следующем слайде покажите результаты симуляции этого способа формировать выборку:

- - сделайте N=10 000 повторов: сформируйте выборку и посчитайте по выборке долю клиентов, купивших продукт
- - постройте гистограмму для полученных N значений, отметьте на оси настоящую (от всех клиентов) долю клиентов, купивших продукт
- - прокомментируйте, есть ли постоянное систематическое завышение или занижение ответа в методе,
- - прокомментируйте, может ли метод метод простой случайной выборки давать большую ошибку, и если может, то часто ли/насколько часто

Красный уровень: подход к решению

Опишите на слайде способ получения стратифицированной случайной выборки

На следующем слайде покажите результаты симуляции этого способа формировать выборку:

- - сделайте N=10 000 повторов: сформируйте выборку и посчитайте по выборке долю клиентов, купивших продукт
- - постройте гистограмму для полученных N значений, отметьте на оси настоящую (от всех клиентов) долю клиентов, купивших продукт
- - прокомментируйте, есть ли постоянное систематическое завышение или занижение ответа в методе,
- - прокомментируйте, может ли метод метод простой случайной выборки давать большую ошибку, и если может, то часто ли/насколько часто

Черный уровень: подход к решению

Сравните на слайде метод простой случайной выборки и метод стратифицированной случайной выборки

- - покажите на одном графике две гистограммы
- - сделайте выводы о точности методов
- - объясните интуитивно, за счет чего стратифицированный метод повышает точность
- - объясните, почему нельзя понять какой из двух подходов лучше по одной выборке простой случайной и одной выборке стратифицированной.

Почему приходится делать симуляцию большого количества повторов обоих методов.

Что и когда нужно сдать?

Что сдавать?

- Jupyter Notebook в Google Colab с расчетами
- PowerPoint презентация со слайдами

Когда сдавать?

- Сроки сдачи указаны в информационной системе

! Сдача кейса – необходимое условие прохождения курса