Тестовое задание для Junior DS

# Цель:

Цель этого тестового задания - понять ваш подход к повседневным бизнес-задачам и проверить ваши навыки в программировании.

Ожидается, что вы будете писать код на языке Python. Можете использовать любые библиотеки. Присылайте свои результаты и визуализацию в отчете, а также код / скрипты для проведения анализа. Убедитесь, что комментариев будет достаточно, чтобы понять, как вы думали, выполняя то или иное задание.

# Задание 1

**Данные:**

1. junior\_d\_meta.pkl

Мета-данные инженера

1. junior\_c\_meta.pkl

Мета-данные email кампании

1. junior\_e\_send.pkl

Данные по отправленным email

1. junior\_e\_open.pkl

Данные по открытым email

**Описание задачи:**

Представим, что компания W выпускает несколько линеек косметической продукции.  
Так же W сотрудничает с косметологами по всей России - рассказывает им о своей продукции и хочет стимулировать их советовать клиентам линейки, выпускаемые W.

Один из способов стимулирования - отправка имейлов с различной информацией о продукции: статьи о безопасности продукта, его эффективности, случаях назначения и тп. (фичи CONTENT\_TYPE, KEY\_MESSAGE)

За рассылки имейлов компания W платит деньги, поэтому в какой-то момент они захотели понять, какую (CONTENT\_TYPE, KEY\_MESSAGE) и когда (в какой день недели и час) информацию нужно отправить каждому отдельному косметологу, чтобы он открыл письмо с наибольшей вероятностью.

**Задача:**

1. Сделать EDA данных с описание того, зачем построены графики и какие выводы из них можно сделать
2. Описать метрику и критерий, по которому она была выбрана
3. Описать схему валидации обучения
4. Построить модель, которая будет предсказывать вероятность открытия косметологом имейла. Интерпретировать модель и ее качество
5. Описать, как полученную модель можно использовать для рекомендации отправляемого контента, какие бизнес-рекомендации можно дать компании W по результатам исследования.

# Описание полей в данных:

**Junior\_d\_meta:**

CITY - город косметолога

REGION1 - информация о гео-локации

REGION2 - информация о гео-локации

REGION3 - информация о гео-локации

CONTACT\_ID - ID инженера

SPECIALTY - специализация косметолога

SEGMENT - сегмент косметолога

**Junior\_c\_meta:**

ID - ID email кампании

BRANDS - бренды косметики, которые участвуют в рекламной кампании

KEY\_MESSAGE - ключевое сообщение имейла

CONTENT\_TYPE - тип контента имейла

CURRENT\_STATUS - статус кампании

**Junior\_e\_open:**

CONTACT\_ID - ID косметолога

ASSET\_ID - ID имейла

ACTIVITY\_TYPE - действие над имейлом

ACTIVITY\_DATE - время совершения действия

**Junior\_e\_send:**

CONTACT\_ID - ID косметолога

ASSET\_ID - ID имейла

ACTIVITY\_TYPE - действие над имейлом

ACTIVITY\_DATE - время совершения действия

CAMPAIGN\_ID - ID email кампании

# Задание 2

1. В таблице T есть поля id (уникальное для каждой строки) и value. Нужно написать SQL-запрос, который бы выводил все id, value которых не уникально.

2. В предположении, что есть все для подключения к BigQuery через Python, нужно написать скрипт, который выгружает таблицу из предыдущего задания в pandas.DataFrame.

# Задание 3

Предположим, есть результат маркетингового эксперимента по замеру эффективности красной/синей кнопки. Результаты известны для сайта и для мобильного приложения.

|  |  |  |
| --- | --- | --- |
| Сайт | Синяя | Красная |
| Нажали | 1000 | 110 |
| Не нажали | 1300 | 250 |
| Соотношение | 0.77 | 0.44 |

|  |  |  |
| --- | --- | --- |
| Мобильное приложение | Синяя | Красная |
| Нажали | 300 | 700 |
| Не нажали | 120 | 330 |
| Соотношение | 2.5 | 2.1 |

При этом общие числа выглядят следующим образом:

|  |  |  |
| --- | --- | --- |
| Сайт + Мобильное приложение | Синяя | Красная |
| Нажали | 1300 | 810 |
| Не нажали | 1420 | 580 |
| Соотношение | 0.92 | 1.4 |

Какие выводы можно сделать по результатам этого исследования?