**Тестовое задание**

**Задание 1. Python и статистика**

Имеется две разные модели, способные определять является ли клиент банка дропом\*. Вероятность обнаружения дропа составляет **90%**, вероятность ошибочно признать клиента дропом составляет **0,1%**. Параметры второй модели: вероятность обнаружения дропа составляет **85%**, вероятность ошибочно признать клиента дропом составляет **2%**. Допустим, что доля дропов среди населения **0,0002**.

*А) Найдите вероятность того, что клиент, определённый моделью как «дроп» окажется порядочным клиентом банка. В чём причина отличия от вероятности исходного ложноположительного результата?*

**ОТВЕТ:**

Для решения данной задачи воспользуемся теоремой и формулой Байеса.

Пусть

Р(М) = 0,0002 – вероятность, что клиент является дропом

Р(Х) = 1 – Р(М) = 0,9998 – вероятность, что клиент хороший

Р(К|М) = 0,9 – вероятность, что первая модель верно определяет дроп

Р(К|Х) = 0,001 – вероятность, что первая модель ошибочно определит хорошего как дроп

Тогда

Р(Х|К) = (Р(К|Х) \* Р(Х)) / (Р(К|М)\*Р(М) + Р(К|Х)\*Р(Х)) =

= 0,001\*0,9998/(0,9\*0,0002+0,001\*0,9998) = 0,8474 или 84,7%

Т.о. вероятность, что клиент, определенный первой моделью как дроп, может оказаться хорошим, составляет примерно 84,7% .

Аналогично можем посчитать для второй модели:

Р(Х|К) = 0,02\*0,9998/(0,85\*0,0002+0,02\*0,9998) = 0,9916 или 99,2%

Причина отличия от вероятности исходного ложноположительного результата в том, что ложноположительный результат – это вероятность, что хороший клиент будет ошибочно классифицирован как дроп. В нашем случае мы знаем что она 0,1%. И это всего лишь процент порядочных клиентов, которые были ошибочно определены как дропы.

В нашем случае учитывая, что дропов очень мало 0,02% от общего числа, и даже если модель ошибается в 0,1% случаев, это не значит, что большинство клиентов, определенных как дропы, являются хорошими. Тут важную роль играет базовая вероятность, что клиент на самом деле является дропом. Поэтому, когда мы комбинируем все вероятности с помощью теоремы Байеса, мы получаем более высокую вероятность того, что клиент действительно хороший, чем просто вероятность ложноположительного результата.

Т.о. разница задается тем, что мы с одной стороны смотрим на вероятность просто отдельной ложноположительной ошибки, а с другой комплексное взаимодействие между вероятностями различных событий.

*Б) Найдите вероятность того, что при повторной проверке клиента уже новой моделью, которая также признает его дропом после признания первой моделью, клиент окажется порядочным.*

**ОТВЕТ:**

Тогда

Р(Х) = 1 – Р(М) = 0,9998 – вероятность, что клиент хороший

Р(К1,К2|Х) = 0,001\*0,02 = 0,00002 – вероятность, что обе модели определяют хорошего как дропа

Р(М) = 0,0002 – вероятность, что клиент является дропом

Р(К1,К2|М) = 0,9 \* 0,85 = 0,765 – вероятность, что обе модели определяют дропа

Получаем

Р = 0,00002 \* 0,9998 / (0,00002 \* 0,9998 + 0,765 \* 0,0002) = 0,1156 или 11,6%

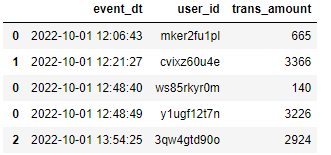
Таким образом, вероятность того, что клиент окажется порядочным после того, как обе модели признали его мошенником, составляет примерно 11,6% .

Смоделировать процесс и подтвердить теоретические рассуждения и полученные вероятности.

*\*Дроп – человек, который за небольшое вознаграждение оформляет на себя обычную банковскую карту, которую затем передаёт мошеннику.*

**Задание 2. SQL**

Имеется набор данных вида (данные во вложении в формате csv):



где event\_dt – дата совершения транзакции, user\_id – идентификатор клиента, trans\_amount – сумма транзакции.

А) Необходимо написать SQL запрос, который выведет для каждой операции по клиенту кумулятивную сумму его расходов в течение суток до момента совершения операции. Например, если клиент совершает операцию в 2022-10-04 12:05:33, то необходимо посчитать сумму транзакций с 2022-10-**03** 12:05:33 по 2022-10-**04** 12:05:33.

**ОТВЕТ:**

WITH num1 AS (

SELECT \*, ((event\_dt) – INTERVAL ‘1 DAY’) as day\_before

FROM table);

SELECT t.event\_dt , t.user\_id, SUM(t.trans\_amount) as total

FROM table AS t, num1

WHERE t.user\_id = num1.user\_id AND

t.event\_dt BETWEEN num1.event\_dt AND num1.day\_before

GROUP BY t.user\_id, t.event\_dt

Б) Найти по паре клиентов на каждую дату, с минимальной и максимальной кумулятивной суммой в этот день.

**ОТВЕТ:**

WITH num1 AS (

SELECT \*, ((event\_dt) – INTERVAL ‘1 DAY’) as day\_before

FROM table),

num2 AS (

SELECT t.event\_dt , t.user\_id, SUM(t.trans\_amount) as total

FROM table AS t, num1

WHERE t.user\_id = num1.user\_id AND

t.event\_dt BETWEEN num1.event\_dt AND num1.day\_before

GROUP BY t.user\_id, t.event\_dt ),

num3 AS (

SELECT event\_dt, MIN(trans\_amount) as minim, MAX(trans\_amount) as maxim

FROM num2

GROUP BY event\_dt);

SELECT num2.event\_dt, num2.user\_id, num2.total

FROM num2, num3

WHERE num2.event\_dt = num3.event\_dt AND

num2.total = num3.minim AND

num2.total = num3.maxim

В) Выполните предыдущие пункты через PySpark DataFrame API.

**Задание 3. Python и анализ данных**

Датасет представляет собой информацию по рынку недвижимости.

**Ссылка на датасет:**

https://www.kaggle.com/datasets/mrdaniilak/russia-real-estate-20182021

Необходимо провести мини-исследование датасета с помощью pandas, matplotlib, seaborn, plotly и других инструментов, а также:

1. Предобработать и почистить данные, заполнить пропуски
2. Определить, какие регионы лидируют по цене. Для каждого региона найти среднюю цену квартиры.
3. Выявить закономерность влияния на цену площади, числа комнат и площади кухни. Влияют ли параметры на цены по регионам? Как можно предсказать цену по входящим параметрам недвижимости?
4. Постройте модель (любую), предсказывающую цену по входящим параметрам недвижимости?

Выявление дополнительных инсайтов будет плюсом. Для каждого пункты необходимо предоставить график с выводами.

**ОТВЕТ:** См. файл во вложении письма.

**Задание 4. Построение DashBoards**

Предположим, что в датасете содержатся данные о покупках в нашем магазине

[Customer Shopping Trends Dataset (kaggle.com)](https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset)

1. Необходимо создать ДБ в DataLens, который бы позволил понимать текущие тренды продаж.

Интерпретация на твое усмотрение, что именно показывать и насколько детально. Приветствуется комментарий, кто в компании будет смотреть на ДБ.

**ОТВЕТ:**

Дашборд построен. <https://datalens.yandex/7dqrtikflsnwu>

Он будет интересен директорам (в том числе директору по продажам), супервайзерам и всем тем, кому интересно знать своего клиента. Дашборд имеет настройки в фильтрах по полу, возрасту и региону продаж.

1. На основании данных подготовить презентацию с рекомендациями, из каких вещей выгоднее сформировать подборку на ярмарку (цвет, размет, тип, пол), чтобы вещи были c проданы с большой вероятностью.

**Задание 5\***

Дан файл в формате .txt с общей информацией о клиентах и истории сайтов, которые они посещали.

Ссылка на данные: <https://disk.yandex.ru/d/1CHYWFCbiS3Haw>

Используя PySpark (или другие библиотеки Python) необходимо извлечь как можно больше ценных инсайтов о пользователях, их поведении, а также визуализировать результаты исследования.