**Описание проекта**

**«Выявление сценариев покупательского поведения**

**среди клиентов бара методом кластеризации»**

**Цель проекта:**

На обезличенных данных выделить сегменты клиентов по их покупательскому поведению. Внутри одного сегмента клиенты должны быть схожи между собой. Клиенты, попавшие в разные сегменты, должны иметь существенные отличия в своих покупках в баре.

Сегментация делается для дальнейшего маркетингового анализа, выделения привлекательных (целевых) сегментов клиентов, разработки адекватного продуктового предложения и позиционирования.

**Задача проекта:**

Методом кластеризации разделить клиентов на группы и описать выявленные сценарии покупательского поведения.

**Данные:**

1. Файл «Depersonalized\_data\_09\_2019.xlsx» содержит информацию о покупках в сети баров в сентябре 2019г. В качестве ID клиента указывается сгенерированное числовое значение. Из демографических данных приведена информация о поле клиента (м/ж). Доходность на 1 клиента приведена в рублях (ARPU), покупки приведены в натуральном выражении (порция, набор).

2. Файл "Directory.xlsx" - для присвоения категорий (групп) покупкам.

**Этапы работы над проектом:**

Созданы 2 ноутбука:

«1.Предобработка и разведочный анализ»

«2.Кластерзация».

**А. Предобработка и разведочный анализ**

1. Цель предобработки: получить дата-фрейм, в котором каждому клиенту соответствует одна строчка с информацией по столбцам. Из матрицы размером 138 021 х 7 получена матрица размером 15 577 х 35.

2. Визуализация и интерпретация существенных данных (разведочный анализ). Ключевым признаком является доходность на 1 клиента в рублях (ARPU).

**B. Кластеризация**

1. **Оцениваем число кластеров** с помощью Метода локтя и Алгоритма DBSCAN. В качестве метрики выбираем Коэф.силуэта, т.к. мы не располагаем размеченными данными. Результат: предварительно рассматриваем версию разбиения на 2-4 кластера.

2. **Визуализируем кластеры**. Для этого стандартизированные наблюдения размещаем на двумерной плоскости с помощью методов снижения размерности PCA и T-SNE.

3. **Находим центры кластеризации** с помощью KMeans.

3.1. Для нестандартизированных данных получаем 3 кластера, кластеры отличаются по кол-ву посещений и ARPU. Данное разбиение отражает 3 разных сценария посещения бара.

3.2. Для стандартизированных данных: неудовлетворительный максимальный коэф.силуэта (0.407) и размытая интерпретация данных.

3.3. Для нормализованных данных: несмотря на неубедительные коэф.силуэта, кластеризация происходит с жёсткой разбивкой по полу (м/ж), что представляет интерес. Получаем 4 кластера, отличающихся по полу (м/ж) и дню недели посещения (сб - для мужчин, пт - для девушек).

4. **Дополнительные расчеты**

4.1. Агломеративная кластеризация

4.2. Иерархическая кластеризация,

4.3. Проверка на данных, которые не содержат аномалий, выявленных алгоритмом DBSCAN.

Подтверждается разбивка на 2-4 кластера.