**Задание**

Необходимо провести речевую аналитику общения клиента и финансового консультанта. Для этого подробно опишите ваши действия для каждого из пунктов предложенного анализа.

**Входные данные:**

Скрипты диалогов финансовых консультантов с клиентами.

В ходе беседы финансовый консультант может предложить клиенту приобрести какой-либо финансовый продукт, рассказать о выгоде нового финансового продукта, назначить встречу для дальнейшего более детального обсуждения продукта, предупредить клиента об истечении срока действия продукта или предложить перезвонить в более удобное время.

**Цель анализа:**

Разделить весь пул клиентов по уровню лояльности («низкая», «средняя», «высокая»).

**Методология анализа:**

1. Удалите выбросы. Какие критерии вы используете для удаления некоторых диалогов?

**Пустые диалоги (бросили трубку), нет дат или названий финансовых продуктов, удалить диалоги где есть ключевые слова о дальнейшем перенаборе.**

1. Придумайте подход для оценки У экзогенным образом. Какие proxy переменные, на ваш взгляд, для этого подходят? Далее в анализе наблюдаемое значение У это и есть выбранное вами proxy переменная.

**Сustomer lifetime value, repeat customer rate (churn rate), net profit score, redemption rate, average purchase amount.**

1. С другой стороны, предположим, что лояльность У это эндогенная переменная, которая определяется набором признаков Х, значение которых определено в ходе диалога. Сформулируйте данный набор признаков, характеризующих лояльность, а также значения, которые они принимают. Чем больше и разнообразнее набор признаков, тем лучше.

**Наличие в диалоге слов, которые характеризуют лояльного/не лояльного пользователя (бинарная переменная), определенная в ходе анализа тональность текста (значение от 1 до 3), близость данного диалога к кластеру лояльных пользователей (расстояние до центра кластера), можно напрямую задавать маркетинговые вопросы.**

1. Выберите форму зависимости и объясните ваш выбор. Опишите используемые метрики качества, а также использованные вами библиотеки, функции и методы анализа.

**Так как мы относим каждый диалог к определенному классу и разметки нет, то это задача кластеризации, поэтому для работы с текстовой кластеризацией подойдут методы word embedding из sklearn (CountVectorizer, TfidfTransformer), genism (word2vec), которые позволят преобразовать исходные данные в вектора и потом уже проводить кластеризацию по нормализованным данным.**

1. Определите границы значений рассчитанной величины лояльности У (если У изначально не категориальная переменная) для каждого уровня («низкая», «средняя», «высокая»). Устойчивы ли они? Опиши способ подбора оптимальной границы.

**Y – категориальная переменная, полученная в результате кластеризации, соответственно при кластеризации нам нужно, чтобы кластеры были максимально далеки друг от друга. Устойчивость можно выявить в результате многократного применения алгоритма на данных: небольшое расхождение результатов будет интерпретироваться как высокая устойчивость.**

1. Опишите способ для упорядочивания выбранного вами набора признаков Х по степени важности для объяснения уровня лояльности У.

**Можно удалять признаки по одному и смотреть как меняется качество классификации, таким образом выявить наиболее существенный.**