Используя Python, составляем витрину данных, которая в обобщённой форме будет отражать историю взаимодействия человека со страховой компанией (содержать информацию о страховом полисе, клиенте, убытках компании)

Все суммы (стоимость полиса, страховую сумму, сумму выплаты по убытку) переводим в доллары США по курсу ЦБ доллара к рублю , полученный по API. (если страховая сумма значится в долларах США изначально, считаем, что курс актуальный) и обавляем эту информацию в виде отдельных столбцов в витрину.

1. Датафрейм содержит информацию о действующих полисах

страхования с атрибутами:

**- contract\_num**— номер контракта;

**- product\_name**— название продукта;

**- client\_id**— уникальный идентификатор страхователя;

**- contract\_status**— статус контракта;

**- currency\_name**— название валюты;

**- duration**— длительность действия полиса;

**- country**— страна проживания клиента;

**- sex**— пол клиента;

**- age**— возраст клиента.

В дополнительных колонках содержится информация о стоимости полиса **(price, price\_usd),** страховой сумме **(insurance\_amount**,**insurance\_amount\_usd)** и сумме выплаты по убытку **(**(**loss\_payout\_amt**,**loss\_payout\_amt\_usd)**  в долларах США по актуальному курсу и в валюте, указанной в графе**currency\_name**.

1. С помощью df.isna().sum() смотрим пропуски, результат следующий :

loss\_payout\_amt 3666

price\_usd 120

insurance\_amount\_usd 120

loss\_payout\_amt\_usd 3668

Других пропусков нет .

Применяя лямбда-функцию, заполняем пропуски в колонках.

1. Проверяем формат данных с помощью df.info() . Результат :

product\_name 3711 non-null object

contract\_status 3711 non-null object

currency\_name 3711 non-null object

country 3711 non-null object

sex 3711 non-null object

C помощью кодировщиков **Label Encoding**и**One-Hot Encoding кодируем эти данные**

1. **Делаем предсказания аномалий с помощью LocalOutlierFactor и IsolationForest**
2. **Получаем итоговый датафрейм с колонкой** anomaly
3. Делаем кластеризацию клиентов , а для этого:

6.1. делим датафрейм на два , в которых раздельно категориалные признаки

и числовые.

6.2. Делаем стандартизацию данных с помощью StandardScaler()

6.3. Применяем метода снижения размерности t-SNE (t-Distributed Stochastic

Neighbor Embedding)

6.4. С помощью метода KMeans делаем кластеризацию на 4 кластера,

количество кластеров выбираем с использованием метода локтя

6.5. С помощью визуализации видно , что два кластера очень похожи и

можно сделать 3 кластера

6.6. Оценка важности признаков с помощью перестановочной важности

с помощью from sklearn.inspection import permutation\_importance

Важные признаки для категориальных признаков contract\_num ( номер контракта) и country (страна проживания клиента) ,

для числовых признаков duration (длительность действия полиса) и

age (возраст клиента)

1. Проводим А/В-тестирование старого и нового подходов к формированию стоимости полиса ВЗР. Первый подход включает в себя традиционную оценку рисков, а второй — использование результатов кластеризации.

Основные влияющие факторы: цена полиса, конверсия в оформления и убыточность.

для контрольной группы средняя цена полиса 51.93$, а наиболее частая цена 47.95$ . для тестовой группы средняя цена полиса 65.34$ , а наиболее частая цена13.7$ . Что говорит у тестовой группы цена полиса более дешевая для большинства клиентов, а у выборочных клиентов цена высокая. Несмотря , что количество клиентов в тестовой группе на 282 меньше, сумма за полисы на 7720$ больше.

конверсия контрольной группы 22.8%

конверсия тестовой группы 26.4%

убыток контрольной группы -21465$

убыток тестовой группы -16560$

По всем вышеперечисленным признакам можно сделать вывод , что методика формирования цены полиса с использованием кластеризации клиентов хорошо себя показала и можно внедрять этот метод.