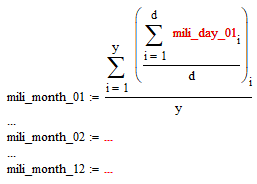
Данные задания предназначены для оценки технических и аналитических навыков кандидата.

1. Требуется **спрогнозировать объем начисленных миль** всем участникам **программы лояльности** в **следующем году**. Источники начислений: **коммерческие полеты**, совершенные участником; совместные **кобрендовые банковские карты**; траты участника на **продукты неавиационных партнеров** (в том числе отели). Приведите факторы, которые необходимо учесть при прогнозировании, и алгоритм оценки их влияния на целевую переменную. Ответ поясните.

Возьмём задачу помесячного прогнозирования для грядущего года. Задачу прогноза для клиента разделим на 3 этапа:

1. **Голая модель**;  
2. **Классовая поправка**;  
3. **Темпоральная** (временная) **поправка**.

1. За **голую модель** возьмём среднемесячное начисление миль клиенту из года в год на каждый месяц:



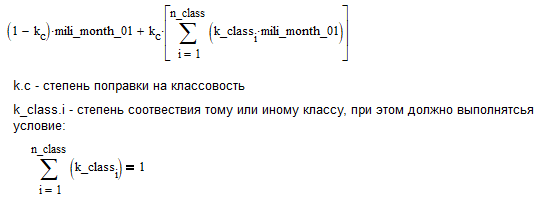
Теперь мы знаем активность для каждого человека на каждый месяц. Глубину поиска по годам стоит вынести в отдельную настройку. Базовый прогноз уже можно составить\*, но имея столько данных, прогноз можно значительно уточнить.

\*Неизвестна степень влияния глобальных изменений на общую тенденцию перелётов. Если она сильна и, допустим, за 10 лет активность перелётов выросла в 2 раза стабильными темпами, то предложенная модель среднего спрогнозирует результаты 5-летней давности. В этом случае такая модель не подходит и тогда необходимо считать ожидаемый прирост, а не среднее. Я за неимением данных не могу сказать, стоит ли делать такие поправки, но предположу, что РФ довольно стабильная страна, **маловероятно**, что она удивляла резким приростом благосостояния граждан и активностью их перелётов. Потому среднее, *скорее всего* подойдёт, даже за 10 лет.

1. Для **классификации** клиента необходимо сформировать классы, а потом отнести клиента по весовым коэффициентам к этим классам и сделать поправку на **классовую принадлежность**. Сформируем классы и критерии к ним:
2. **Отпускник**. Анализируем данные **неавиационных партнёров**, вычленяем клиентов, тратившихся на отели, отсекаем по пороговому значению (а вот, как устанавливать порог – воля аналитика, возьмём, например, не менее 0,6836 траты на отель в год)
3. **Бизнесмен**. Выборка людей, покупающих преимущественно бизнес-класс, допустим, 80%
4. **Медиум**. Менее 80% бизнес-класс перелётов
5. **Эконом**. Менее 20% бизнес-класс перелётов
6. **Командировочник**. Частые перелёты (насколько?)
7. **Попаданец**. Крайне редкие полёты
8. **Патриот**. Преимущественно перелёты внутри РФ, допустим, 80%
9. **Ренегат**. Преимущественно перелёты за пределами РФ, допустим, 80%
10. **Половая принадлежность**
11. **Территориальная принадлежность**
12. **Партнёр**. Владелец ко-брендовой карты, от них ожидается повышенная лояльность.
13. …

Вопрос подбора количества классов и вопрос включения/исключения их в формировании модели решаются путём тестирования.

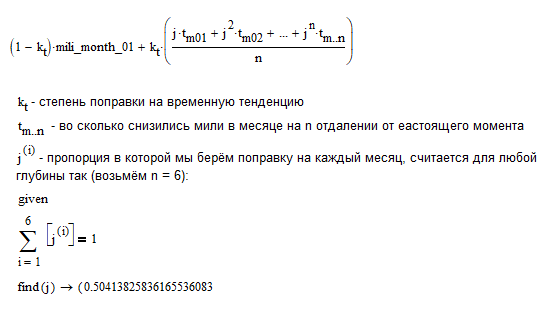
Далее необходимо по уже упомянутому принципу сформировать модели «среднего» представителя класса. Определить принадлежность (или степень принадлежности) голой модели к классам и сделать поправку по формуле:



Чтобы вышеупомянутое условие выполнялось, после определения соответствия на классы необходимо провести нормировку к 1. Коэффициенты соответствия классу необходимо вычислять через свертку множеств или близость к критериям.

Теперь мы усреднили **голую модель** по её **классовой принадлежности**.

1. **Темпоральная поправка** вводится для учёта последних тенденций (пример – пандемия). Рассчитывается общее снижение объёмов начисления миль для всех для последних n месяцев по соотношению к среднему за предыдущие года. Далее каждый последующий прогнозируемый месяц модели множится на темпоральный коэффициент по принципу отдаления от настоящего момента, приведу пример для первого месяца месяца:



Нумерация поправок соответствует не номеру месяца, а отдалению от настоящего момента, если сейчас июль, то темпоральная поправка на август от июня будет **tm01**, от мая **tm02**, от апреля **tm03** и тд. Далее рассматривается поправка на сентябрь вся нумерация смещается.

В моём описании много ключевых коэффициентов, подбирать которые должен аналитик. Эти коэффициенты необходимо вынести в отдельное окно настроек, это позволит аналитику менять степень влияния факторов. Например, для клиента с большим количеством перелётов не обязательно делать поправку на класс, т.к. его базовая модель и без того будет точной. Или если за последний год не было сильных скачков активности, то темпоральная поправка не окажет существенного влияния.

1. Используя диалект **Oracle SQL**, сформировать запрос к хранилищу данных, позволяющий получить **номера участников** в программе лояльности, **фактически** **потративших на авиабилет более 50 000 миль** **за 2020 год** и имевших **до совершения** **первой** **полетной** **транзакции в 2020 году более 40 000 миль** **на счету**. Требуется вывести **номер участника** в программе, **к-во транзакций списания на билеты** и **среднее количество миль, списанных на билеты, в 2020 году**.

Структура необходимых таблиц в хранилище приведена ниже.

Таблица с участниками:

|  |  |
| --- | --- |
| **Client id** | **Client number** |
| Client id 1 | Client number 1 |
| Client id 2 | Client number 2 |
| Client id 3 | Client number 3 |
| … | … |

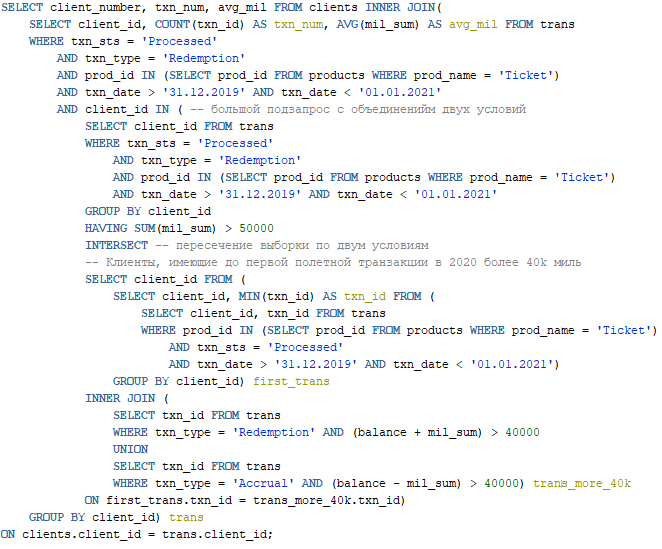
Таблица с транзакциями:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Client id** | **Txn ID** | **Txn\_date** | **Txn\_type** | **TXN\_sts** | **Product id** | **Miles\_sum** | **Client mile balance\_after txn** |
| Client id 1 | Txn ID 1 | 01.01.2019 | Accrual | Cancelled | Product type id 1 | Miles\_sum 1 | Balance\_after txn 1 |
| Client id 2 | Txn ID 2 | 27.03.2019 | Redemption | Processed | Product type id 2 | Miles\_sum 1 | Balance\_after txn 2 |
| Client id 3 | Txn ID 3 |  | Redemption | Cancelled | Product type id 3 | Miles\_sum1 | Balance\_after txn 3 |
| … | … | 12.11.2018 | … | … | … | … | … |

Таблица с продуктами:

|  |  |
| --- | --- |
| **Product id** | **Product\_nm** |
| Product id 1 | Ticket |
| Product id 2 | Bank card |
| Product id 3 | Hotel |
| … | … |

Выполним запрос\*



SQL-файлы создания тестовой таблицы и запроса высылаю вместе с тестовым заданием

\*вы время выполнения задания возникли сложности трактовки текстового описания при формализации их в SQL синтаксис, например, не всегда однозначно ясно, что делать с отменёнными запросами. Выполнил по своему разумению.

1. Требуется разработать сегмент для рассылки по **стимулированию совершения полетов в определенные города России в определенный временной интервал (например, декабрь-январь)** путем предоставления **промокода со скидкой** на покупку билета. Предложите критерии, по которым будет осуществлен отбор участников программы в сегмент так, чтобы коммуникация была получена наиболее потенциально заинтересованными в предложении, а также алгоритм оценки эффективности коммуникации.

**Критерии подбора участников программы**

1. Те клиенты, которые часто совершают перелёты **в** данные города\*
2. Те клиенты, которые совершают перелёты **из** данных городов в близкие промежутки времени, ведь им понадобится вернуться
3. Клиенты, совершающие поездки в различные (не одни и те же!) города РФ в данные периоды времени
4. Сторонние источники информации о клиентах, по которым есть возможность вычленить заинтересованных в поездках клиентов (например, мы знаем, что у клиента живут в городе родственники, знаем, что в город развит определённый вид спортивного туризма и клиент совершает покупки сопутствующего инвентаря и тд..)
5. Я сперва хотел предложить исключить из выборки клиентов, которые летают, приобретая билеты в рамках покупки путёвки, т.к. покупают они путёвку с включённым билетом, а не билет, и маловероятно, что им нужны будут билеты. Но решил, что стоит исключать не всех. Необходимо оставить из них тех клиентов, кто очень часто путешествует по РФ с целью туризм, т.к. есть вероятность, что они решат совершить путешествие не через путёвку, а самостоятельно.

\* если выборка клиентов по 1 пункту значительно больше целевой, можно ограничить их дополнительным временным условием, или делать критерии частоты полётов более высокими, или включить в выборку тех клиентов, кто до этого откликался на подобные акции.

**Алгоритм оценки эффективности коммуникации**

Дать качественную оценку клиентов, которые воспользовались промокодом.

Дать качественную оценку клиентов, которые промокод получили, совершили поездку, но промокодом не воспользовались. Этот сегмент – явно недостигшая цели коммуникация.

Оценить эффективность коммуникации через оценку разности среднемесячного показателя перелётов за предыдущие года к нынешнему среди двух выборок: среди всех клиентов, среди клиентов, выбранных для рассылки.

Проверить «утечки» промокодов. Какие клиенты не были в рассылке, но получили промокод (от других клиентов или с сайтов с раздачами) и воспользовались им. Возможно, эти данные помогут оптимизировать критерии в дальнейшем.

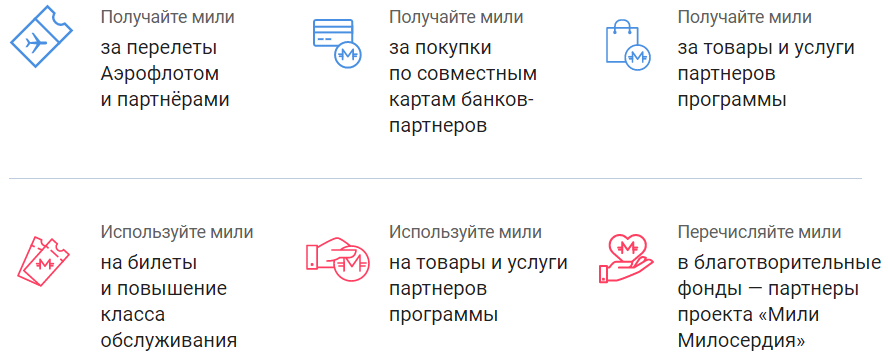
1. Необходимо построить модель, предсказывающую **количество купленных авиабилетов** в зависимости **от** **ряда переменных**. Перечислите эти переменные и обоснуйте, почему они должны быть включены в модель.

Переменные, которые стоит учитывать в модели:

1. **Время, сезонность**. Позволит с помощью модели определять рост и спад продаж в разные сезоны.
2. **Средняя стоимость билетов** **по категориям** в долларах с учётом инфляции. Стоимость билетов может зависеть от множества факторов, но очевидно, что количество покупаемых билетов будет меньше при большей стоимости при идентичных условиях
3. Соотношения цены на **перелёт эконом** класса **к цене поездки на поезде**
4. Соотношение цены билетов **Аэрофлота** **к билетам конкурентов**
5. Открытые **экономические данные**. Медианный или средний доход, думаю, будет наиболее весом
6. **Международный климат**. Закрытие границ для перелётов разных стран или сильное изменение привлекательности стран для перелётов
7. Количество владельцев **карт лояльности|ко-брендовых карт**. Такие клиенты будут с большей вероятностью пользоваться услугами аэрофлота

Степень влияния того или иного фактора в отдельности необходимо высчитывать, применяя математические методы из статистики. В данном случае применим регрессионный анализ.

1. Приведите пример **алгоритма сегментирования** **базы участников** в зависимости от **степени активности в программе лояльности**. Какой показатель лучше взять в качестве меры активности и как его измерить?



Для оценки степени активности и дальнейшем сегментировании клиентов я предлагаю использовать показатель **оборота (сумму прихода и расхода) миль за месяц**:

, где

– приход миль

– расход миль

– оборот миль

– количество месяцев в программе, может быть дробным, но не менее единицы

При основной сегментации стоит отсекать следующих участников, т.к. их показатели не презентативны, вынести их в отдельные сегмент:

1. **Новичков** - участвующих в программе менее месяца
2. **Разовиков** - совершивших **менее двух** актов пополнения миль (или n - на усмотрение аналитика)
3. **Мертвых душ** - не совершавших движений по программе более t месяцев

Упомянутые 3 сегмента в другие выборки включать не стоит.

Сегментацию по *k* предлагаю следующую:

1. **Крайне редко пользующиеся**
2. **Малоактивные**
3. **Средней активности**
4. **Активные**
5. **Очень активные**

Конкретные цифры отсечек должны быть выбраны из анализа функции количества пользователей от активности. Можно, конечно, взять стандартные 20/80

Помимо сегментации по параметру степени активности *k*, стоит дополнительно сегментировать пользователей по следующим признакам:

1. **Плюшкины** – копящие, но не тратящие мили, им следует напомнить о преимуществах программы
2. **Оптовики** – имеющие редкие по частоте, но большие по объёму притоки миль (туда попадут, например те, кто ездит в Америку пару раз в год и никуда больше)
3. **Бедные, но верные** – имеющие низкий k, но пользующиеся программой без прерываний (туда попадут, например те, кто стабильно летает экономом Москва-Питер)