Создание вероятностной модели покупки товара используя SQL-код

Цель: увеличить доход магазина за счет дополнительной продажи горных велосипедов существующим и новым клиентам.

Определить целевую аудиторию которая с максимальной вероятностью положительно отреагирует на предложение о покупке горного велосипеда.

Используя концепцию Байесовской условной вероятности определить клиентов магазина, которые не покупали в нашем магазине горные велосипеды, но на 80-100% по своим характеристикам (пол, возраст, доход и прочие исследованные данные) соответствуют покупателям ранее купившим горный велосипед.

Для достижения поставленной цели: информировать предложением о покупке следующие группы клиентов с идентичными характеристиками:

* существующих клиентов магазина не купивших ранее горный велосипед;
* вновь регистрируемых клиентов в базе данных магазина;
* потенциальных покупателей, не являющихся клиентами магазина, но по имеющимся данным соответствующих нашей выборке.

В базе данных магазина накоплен большой объем поведения 18’484 покупателей (массивы DimCustomer, DimGeography учебной базы MS AdventureWorksDW2014) и их покупках (массивы FactInternetSales, DimProduct).

Предполагая стабильность прошлых факторов в будущем можно экстраполировать поведение текущих покупателей на будущих: надо взять большой массив в прошлых данных на основе них построить вероятностную модель с выявленными условиями.

Через эту модель проанализировать: текущих клиентов (которые еще не купили исследуемый товар) + всех вновь регистрируемых у нас клиентов + потенциальных клиентов.

Используя SQL – обработка данных быстрее за счет произведения расчетов в месте хранения данных.

Шаги реализации:

В запросе №1 производится набор характеристик для 18’484 покупателей (семейное положение=MaritalStatus, возраст=age, доход=<...>, гендерность, наличие детей, место проживания и пр.) с агрегацией в подзапросе факта (1 or 0) покупки горного велосипеда (ProductSubcategoryKey = 1).

***** запрос №1*****

```
create view PastEvents
```

```
        AS
    select  CustomerKey, MaritalStatus, ROUND(DATEDIFF(year, BirthDate, getdate()), -1)
    as age, Gender, YearlyIncome, TotalChildren, NumberCarsOwned as Cars,
    HouseOwnerFlag, EnglishEducation as Education, EnglishOccupation as Occupation,
    g.EnglishCountryRegionName as Country,
                (
                select sign(COUNT(*))
                from FactInternetSales as s inner join DimProduct as p
                            on s.ProductKey = p.ProductKey
                where ProductSubcategoryKey = 1
                        and s.CustomerKey = c.CustomerKey
                ) as mountinBiker
    from DimCustomer as C inner join DimGeography as G
        on c.GeographyKey = g.GeographyKey
*****
```

В запросе №2: осуществляется разделение на два столбца

```
**** Запрос №2 ******

    select *,
        case mountinBiker
            when 1 then 1
        end as yes,
        case mountinBiker
            when 0 then 0
        end as no
    from PastEvents
*****
```

В запросе №3: определен результат в кластерах по возрасту клиентов: промежуточный вывод: 24% покупателей в возрасте 50 лет купили горный велосипед. Дальнейшие действия : прибавляя характеристики найти в значимых кластерах долю покупки более 80%.

Результаты | Сообщения

	Age	купили	не купили
1	40	0.2003186	0.7996814
2	50	0.2412093	0.7587907
3	60	0.2377009	0.7622991
4	70	0.208206	0.791794
5	80	0.1754587	0.8245413
6	90	0.09876543	0.9012346
7	100	0.04545455	0.9545454

```
**** Запрос №3****
    with History
        AS (
          select *,
```

```
                    case mountinBiker
                        when 1 then 1
                    end as yes,
                    case mountinBiker
                        when 0 then 0
                    end as no
                from PastEvents
                )
    select Age,
                CAST(Count(yes) as Real)/count(*) as [купили],
                CAST(Count(no) as Real)/count(*) as [не купили]
    from History
    group by Age
    order by Age
```

*****

В запросе №4: при сочетании 10-ти характеристик собраны кластеры с сортировкой по убыванию. В первой строке таблицы: в группе из 82 человек (столбец Total) с указанными характеристиками: 37,8% купили горный велосипед.

Результаты | Сообщения

	MaritalStatus	Age	Gender	YearlyIncome	TotalChildren	Cars	HouseOwnerFlag	Education	Occupation	Country	купили	не купили	Total
1	M	60	M	40000.00	1	1	1	Partial College	Clerical	United States	0.3780488	0.6219512	82
2	M	60	F	40000.00	1	1	1	Partial College	Clerical	United States	0.3768116	0.6231884	69
3	M	60	F	60000.00	1	1	1	Partial College	Skilled Manual	United States	0.2982456	0.7017544	57
4	S	50	M	70000.00	0	1	0	Bachelors	Professional	Australia	0.2105263	0.7894737	57
5	S	50	F	70000.00	0	1	0	Bachelors	Professional	Australia	0.2909091	0.7090909	55
6	M	50	F	40000.00	1	0	1	Bachelors	Skilled Manual	United Kingdom	0.09803922	0.9019608	51
7	S	40	F	40000.00	0	2	0	High School	Skilled Manual	United States	0.18	0.82	50
8	M	40	F	40000.00	0	2	1	High School	Skilled Manual	United States	0.1489362	0.8510638	47
9	M	50	M	40000.00	1	0	1	Bachelors	Skilled Manual	United Kingdom	0.1590909	0.8409091	44
10	M	40	M	40000.00	0	2	1	High School	Skilled Manual	United States	0.2564103	0.7435898	39

**** Запрос №4 *****

```
    with History
            AS (
              select *,
                    case mountinBiker
                        when 1 then 1
                    end as yes,
                    case mountinBiker
                        when 0 then 0
                    end as no
                from PastEvents
                )
    select MaritalStatus, Age, Gender, YearlyIncome, TotalChildren, Cars,
    HouseOwnerFlag, Education, Occupation, Country,
                CAST(Count(yes) as Real)/count(*) as [купили],
                CAST(Count(no) as Real)/count(*) as [не купили],
                Count(*) as Total -- определение значимости кластера для анализа
    from History
```

```
group by MaritalStatus, Age, Gender, YearlyIncome, TotalChildren, Cars,
HouseOwnerFlag, Education, Occupation,   Country

order by Total desc
```

******

В запросе №5 выделены в том числе 22 кластера (с группами свыше 10 человек) где все 100% людей купили горные велосипеды. В остальных кластерах выборки доля купивших горные велосипеды от 80 до 99%.

****** Запрос №5 ******

```
with History
        AS (
          select *,
                        case mountinBiker
                                when 1 then 1
                        end as yes,
                        case mountinBiker
                                when 0 then 0
                        end as no
                from PastEvents
                )
select MaritalStatus, Age, Gender, YearlyIncome, TotalChildren, Cars,
HouseOwnerFlag, Education, Occupation, Country,
                CAST(Count(yes) as Real)/count(*) as [купили],
                CAST(Count(no) as Real)/count(*) as [не купили],
                Count(*) as Total -- определение значимости кластера для анализа
from History
group by MaritalStatus, Age, Gender, YearlyIncome, TotalChildren, Cars,
HouseOwnerFlag, Education, Occupation, Country
                with cube
having Count(*)> 10
        and CAST(Count(yes) as Real)/count(*) > 0.8
order by [купили] desc
```

****

В запросе №6 формирование таблицы результатов: into [покупатели горных велосипедов].

****** Запрос №6 ******

```
with History
        AS (
          select *,
                        case mountinBiker
                                when 1 then 1
```

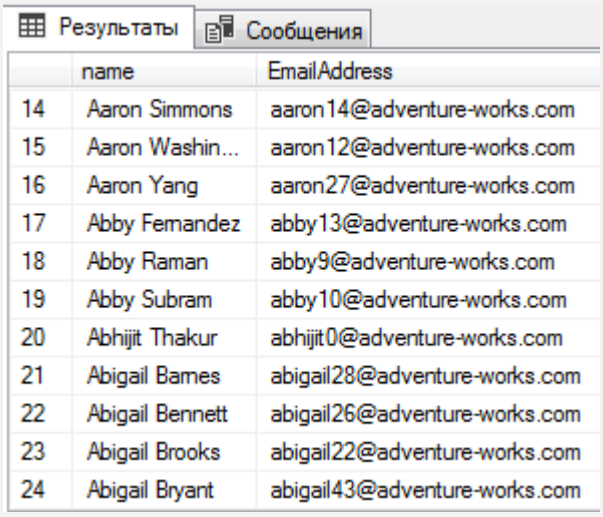
```
                    end as yes,
                    case mountinBiker
                        when 0 then 0
                    end as no
                from PastEvents
                )
    select MaritalStatus, Age, Gender, YearlyIncome, TotalChildren, Cars,
    HouseOwnerFlag, Education, Occupation, Country
    into    [покупатели горных велосипедов]
    from History
    group by MaritalStatus, Age, Gender, YearlyIncome, TotalChildren, Cars,
    HouseOwnerFlag, Education, Occupation, Country
                with cube
    having Count(*)> 10
        and CAST(Count(yes) as Real)/count(*) > 0.8
******
```

Проверку модели произведу на текущей таблице клиентов (без присоединения к ней таблицы продаж). В запросе №7 формирование таблицы результатов для отправки сообщений.

****** Запрос №7 ******

```
    select distinct
                C.FirstName + ' ' + c.LastName as name,
                C.EmailAddress
    from [AdventureWorksDW2014].[dbo].[DimCustomer] as C inner join [покупатели
    горных велосипедов] as MB
        on (c.Gender = mb.gender or mb.gender is null)
        and (c.NumberCarsOwned = mb.Cars or mb.cars is null)
        and (ROUND(DATEDIFF(year, c.BirthDate, getdate()), -1) = mb.age or mb.age is
    null)
        and (c.YearlyIncome = mb.YearlyIncome or mb.YearlyIncome is null)
        and (c.MaritalStatus = mb.MaritalStatus or mb.MaritalStatus is null)
        and (c.Gender = mb.Gender or mb.Gender is null)
        and (c.TotalChildren = mb.TotalChildren or mb.TotalChildren is null)
        and (c.HouseOwnerFlag = mb.HouseOwnerFlag or mb.HouseOwnerFlag is null)
        and (c.EnglishEducation = mb.Education or mb.Education is null)
        and (c.EnglishOccupation = mb.Occupation or mb.Occupation is null)
        and (G.EnglishCountryRegionName = mb.Country or mb.Country is null)
******
```

Результатом является база имен и адресов почты для отправки сообщений.



Результаты | Сообщения

	name	EmailAddress
14	Aaron Simmons	aaron14@adventure-works.com
15	Aaron Washin...	aaron12@adventure-works.com
16	Aaron Yang	aaron27@adventure-works.com
17	Abby Fernandez	abby13@adventure-works.com
18	Abby Raman	abby9@adventure-works.com
19	Abby Subram	abby10@adventure-works.com
20	Abhijit Thakur	abhijit0@adventure-works.com
21	Abigail Barnes	abigail28@adventure-works.com
22	Abigail Bennett	abigail26@adventure-works.com
23	Abigail Brooks	abigail22@adventure-works.com
24	Abigail Bryant	abigail43@adventure-works.com

Точность достижения цели можно повысить:

- Учесть сезонность покупки в модели;
- Дату покупки с анализом диапазонов от начала сезона и в сезон;
- Уменьшением базы рассылки клиентам магазина (по нашим данным не имеющим горный велосипед) на показатель покупки аксессуаров и/или запчастей к горному велосипеду, то-есть данным косвенно подтверждающим наличие горного велосипеда, купленного в другом магазине

Евгений Крылов
+7 931 300 1456
krylove77@gmail.com

июнь 2022