АНАЛИЗ ТРАНЗАКЦИОННОЙ БАЗЫ ДАННЫХ. АССОЦИАТИВНЫЕ ПРАВИЛА

Транзакция – это множество единовременно совершенных покупок, оплаченных одним чеком. *Транзакционная или операционная база данных* представляет собой двумерную таблицу, которая состоит из номера транзакции (TID) и перечня покупок, приобретенных во время этой транзакции.

TID – это уникальный идентификатор, определяющий каждую сделку или транзакцию.

На рисунке 62 приведен фрагмент транзакционной базы данных покупок в супермаркете. Это двумерная таблица, состоящая из номеров транзакций (TID) и перечня покупок, приобретенных во время этих транзакции. Наличие единицы в ячейке таблицы говорит о том, что в момент транзакции (заголовок строки таблицы) был куплен товар (заголовок столбца таблицы). Количество единиц в одной строке таблицы характеризует количество единовременно приобретенных категорий товаров.

	A	B	C	D	E	F	G	H	I	J
1	TID	молоко	помидоры	конфеты	чай	огурцы	кофе	масло	сметана	хлеб
2	10		1		1	1		1	1	
3	11	1		1			1		1	1
4	12		1			1			1	
5	13	1		1	1		1	1		1
6	14				1	1		1		1
7	15			1		1	1	1	1	1
8	16		1	1	1	1			1	1
9	17				1			1		1
10	18			1		1	1	1		1
11	19	1			1	1	1			1
12	20	1	1			1	1	1	1	1
13	21	1			1			1		1
14	22			1	1	1		1		1
15	23	1		1	1	1		1		
16	24	1	1		1	1	1			1
17	25			1	1		1	1	1	1
18	26	1		1	1	1		1		1
19	27	1		1		1	1			
20	28				1			1	1	1
21	29		1	1	1	1				
22	30				1			1		1
23	31	1		1	1	1	1	1	1	1
24	32	1		1		1		1		
25	33	1	1	1	1		1	1		1
26	34	1		1	1	1		1	1	1

Рис. 62. Фрагмент транзакционной базы данных

Ассоциация – это поиск правил и закономерностей между связанными событиями в наборе данных.

В общем виде ассоциативное правило можно описать: «Из события *B* следует событие *C*» или «Кто купил товар *B*, также купит товар *C*».

Основными характеристиками ассоциативного правила являются *поддержка*, *достоверность* и *лифт* правила.

Поддержка (SUPP) – количество или процент транзакций из всего множества, содержащих определенный набор данных.

Достоверность правила (CONF) – мера точности правила, характеризующая, например, какой процент транзакций из всего множества, содержащих набор элементов b, также содержат набор элементов c.

$$\mathrm{CONF}(b \to c) = \frac{\mathrm{SUPP}(bc)}{\mathrm{SUPP}(b)}$$

Лифт (LIFT) – оценка независимости событий друг от друга (оценка значимости правил).

$$LIFT(b \to c) = \frac{SUPP(bc)}{SUPP(b) \times SUPP(c)}$$

$LIFT = 1$ – событие в левой части независимо от события в правой части. Если два события независимы, то никакого правила не существует.

$LIFT < 1$ – наличие одного события имеет отрицательный эффект на возникновение другого. События не могут сосуществовать.

$LIFT > 1$ – указывает степень связи между событиями. Величина, на которую лифт, собственно, больше этой самой единицы, и покажет нам «силу» правила. Чем больше единицы, тем лучше.

Если $LIFT(b \to c) = 1{,}43$ – это означает, что правило: «Из покупки набора элементов b следует покупка набора элементов c» – на 43% мощнее правил, о том, что набор b или набор c просто покупают.

Задача 1

На основании данных транзакционной базы данных покупок в супермаркете (см. рис. 62) найти три наиболее часто покупаемых товара и проверить ассоциативное правило: «Из покупки самого популярного товара следует покупка двух других найденных популярных товаров».

Ход решения

- Для каждого наименования товара рассчитать, какой процент транзакций из всего множества содержат данный товар. На основании этого определить три наиболее популярных товара.
- Исследовать ассоциативное правило: «Из покупки самого популярного товара следует покупка двух других найденных популярных товаров», рассчитав поддержку, достоверность и лифт этого правила.

Решение

1. Создадим лист Excel с именем «База данных» и заполним его ячейки значениями из таблицы (см. рис. 62).

2. Определим три наиболее популярных товара в этой базе данных, вычислив для каждого наименования товара, какой процент транзакций из всего множества содержат данный товар, т.е. рассчитаем поддержку правил о том, что данные товары покупают.

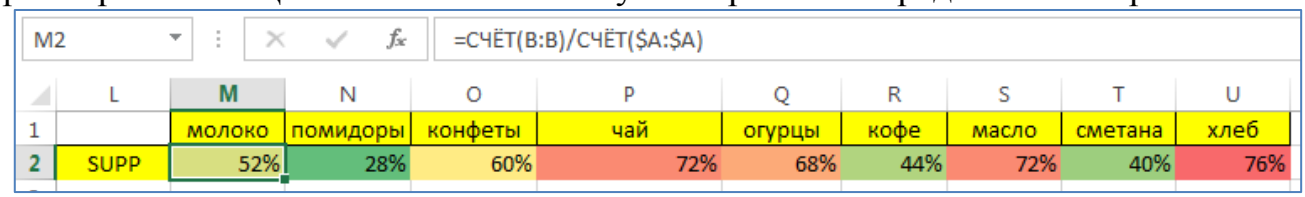
2.1. Для этого в ячейки M1:U1 скопируем наименования товаров из ячеек B1:J1.

2.2. В ячейку M2 введем формулу расчета SUPP(молоко), рассчитав тем самым, какой процент транзакций из всей базы данных содержат товар «молоко»:

=СЧЁТ(B:B)/СЧЁТ($A:$A)

2.3. Распространим введенную формулу правее по строке, заполнив ей диапазон ячеек M2:U2.

2.4. Установим для диапазона ячеек M2:U2 процентный формат и условное форматирование «Цветовые шкалы». Результат расчетов представлен на рис. 63.



	L	M	N	O	P	Q	R	S	T	U
1		молоко	помидоры	конфеты	чай	огурцы	кофе	масло	сметана	хлеб
2	SUPP	52%	28%	60%	72%	68%	44%	72%	40%	76%

Рис. 63. Процентная характеристика популярности товаров

2.5. Из расчетов (рис. 63) видно, что чаще всего покупают «*хлеб*». Следующими по популярности являются товары «*чай*» и «*масло*».

3. Исследуем ассоциативное правило:

«Если купили *хлеб*, то также купят *чай* и *масло*».

3.1. Рассчитаем поддержку этого правила: SUPP(хлеб чай масло). Для этого в ячейку M6 введем формулу:

=СУММПРОИЗВ(J:J;E:E;H:H)/СЧЁТ(A:A).

$\mathrm{SUPP}(\text{хлеб чай масло}) = 48\%$ означает, что почти половина всех транзакций в базе данных содержат этот набор товаров.

3.2. Рассчитаем достоверность правила:

$$\mathrm{CONF}(\text{хлеб} \rightarrow \text{чай масло}) = \frac{\mathrm{SUPP}(\text{хлеб чай масло})}{\mathrm{SUPP}(\text{хлеб})}.$$

Для этого в ячейку M7 введем формулу:

=M6/U2.

$\mathrm{CONF}(\text{хлеб} \rightarrow \text{чай масло}) = 63\%$ характеризует довольно высокую точность этого правила. Более половины транзакций, содержащих «*хлеб*», также содержат набор товаров «*чай*» и «*масло*».

3.3. Рассчитаем лифт правила:

$$LIFT(\text{хлеб} \rightarrow \text{чай масло}) = \frac{SUPP(\text{хлеб чай масло})}{SUPP(\text{хлеб}) \times SUPP(\text{чай}) \times SUPP(\text{масло})}$$

Для этого в ячейку M7 введем формулу:

=M6/(U2*P2*S2).

$LIFT(\text{хлеб} \rightarrow \text{чай масло}) = 1{,}218$ означает, что правило: «Если купили *хлеб*, то также купят *чай* и *масло*» – на 22% мощнее правил, о том, что «*хлеб*», «*чай*» и «*масло*» просто покупают по отдельности.

3.4. Для ячеек M6 и M7 установим процентный формат и оформим наши расчеты так, как это показано на рисунке 64.

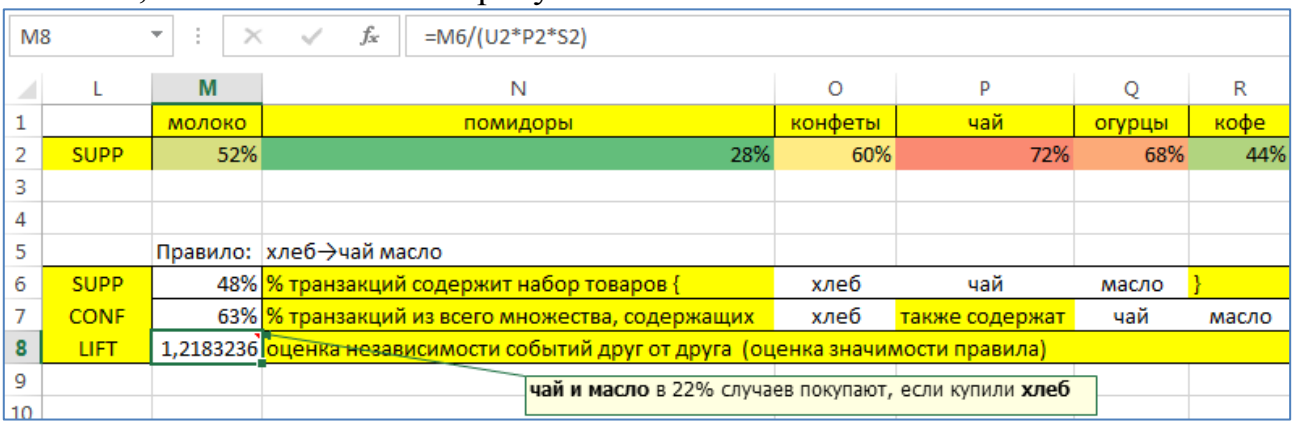


Рис. 64. Итоговый результат построения модели

Задача 2

На основании данных транзакционной базы данных покупок в супермаркете (см. рис. 62) исследовать ассоциативные правила:

- Если купили *кофе*, то также купят *молоко* и *конфеты*;
- Если купили *кофе*, то также купят *помидоры* и *сметану*.

Расчеты оформить по образцу, представленному на рисунке 65.

Что означают полученные значения поддержки, достоверности и лифта правил?

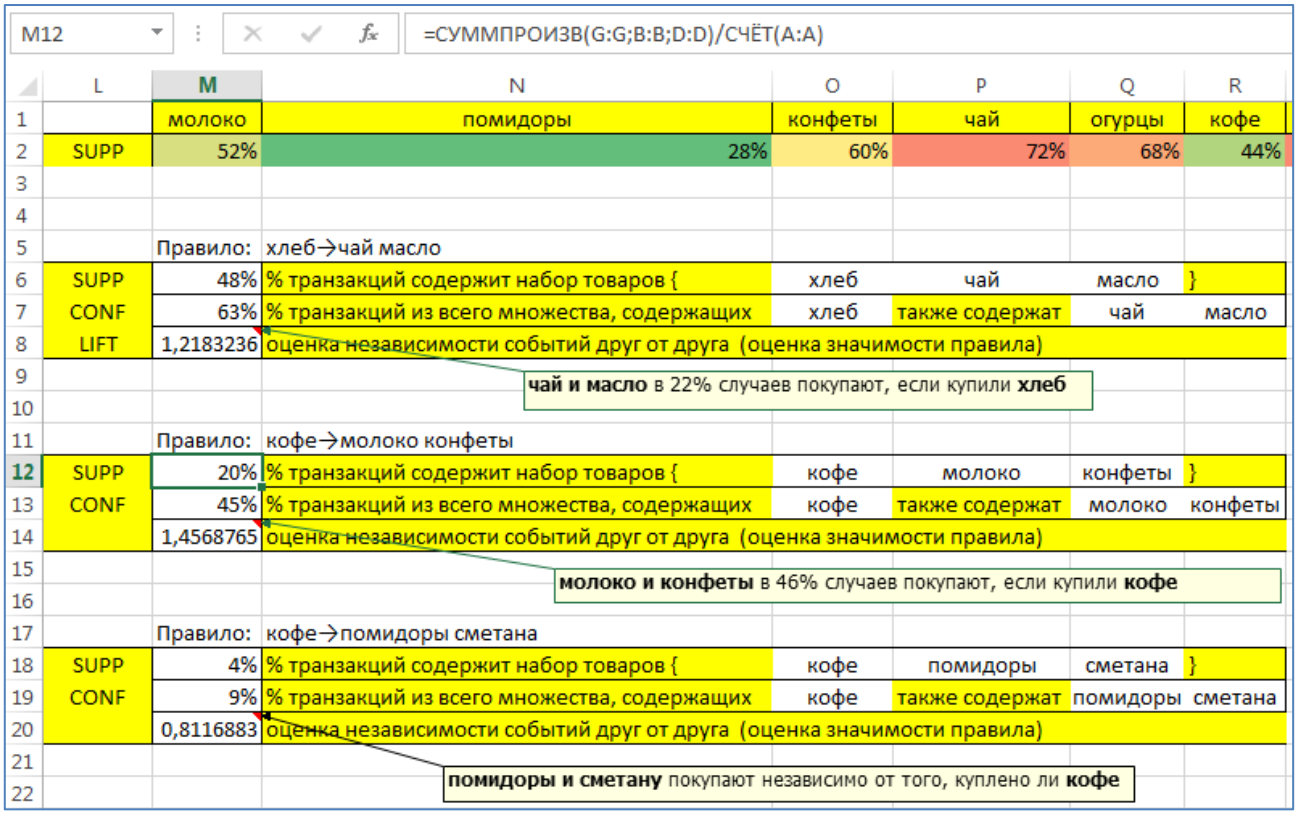


Рис. 65. Итоговый результат построения модели для задачи №2