

**СОВМЕСТНОЕ ИСПОЛЬЗОВАНИЕ ПАКЕТА
MICROSOFT OFFICE EXCEL 2007 и СЛУЖБ
ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ SQL
SERVER ANALYSIS SERVICES**

Афанасьева С.В.

Data Mining (Интеллектуальный анализ данных) - это технология выявления скрытых взаимосвязей внутри больших баз данных.

Является службой Microsoft SQL Server 2005 (2008) Analysis Services

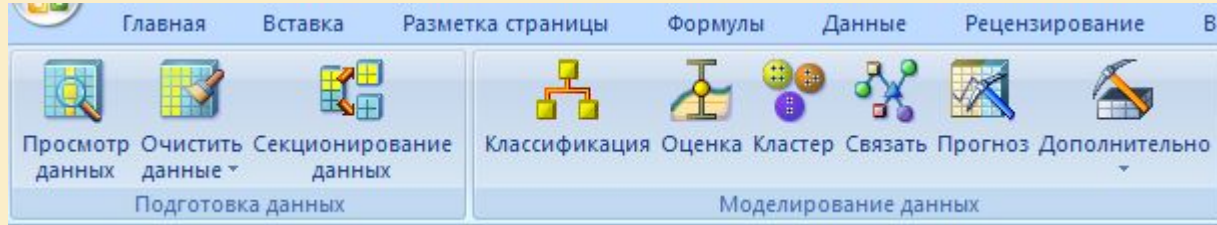
СЛУЖБЫ MICROSOFT SQL SERVER 2005 ANALYSIS SERVICES (SSAS) СОДЕРЖАТ :

- Алгоритм дерева принятия решений
- Алгоритм кластеризации
- Упрощенный алгоритм Байеса
- Алгоритм взаимосвязей
- Алгоритм кластеризации последовательностей
- Алгоритм временных рядов
- Алгоритм нейронной сети (службы SSAS)
- Алгоритм логистической регрессии
- Алгоритм линейной регрессии

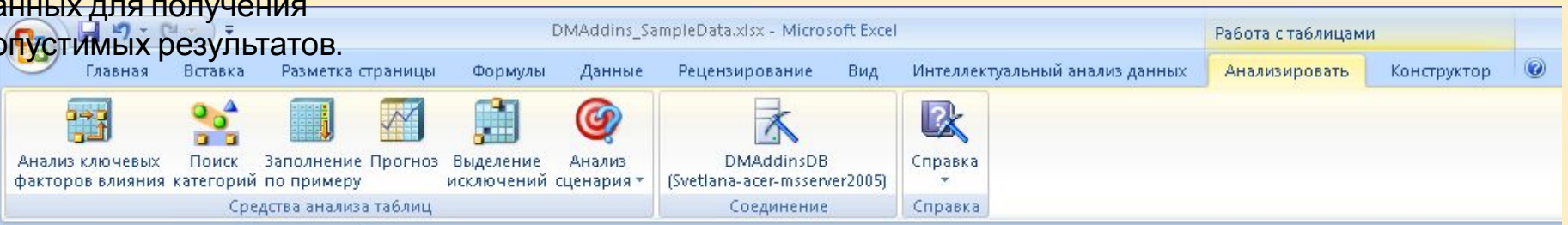
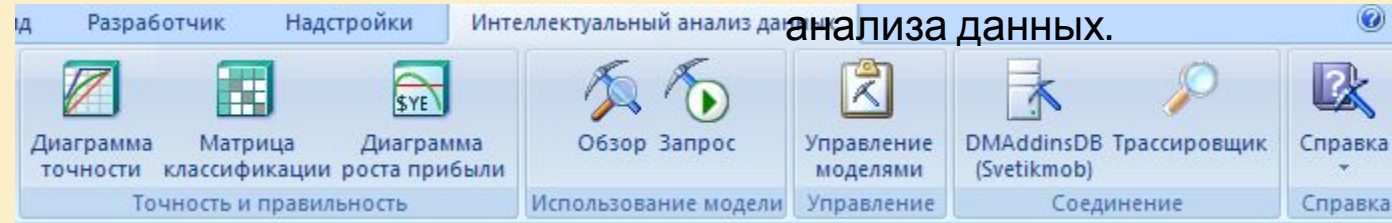
НАДСТРОЙКИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ SQL SERVER 2005 ДЛЯ OFFICE 2007

- ❖ выявлять закономерности и тренды, существующие в сложных данных,
- ❖ отображать такие закономерности в диаграммах и интерактивных средствах просмотра
- ❖ формировать цветные сводные отчеты для презентаций и бизнес-аналитики.
- ❖ анализировать корреляции и формировать прогнозы для данных, хранящихся в таблицах Microsoft Office Excel, или создавать и изменять модели интеллектуального анализа данных, хранящихся в экземпляре Analysis Services

НАДСТРОЙКИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В OFFICE 2007



Алгоритм интеллектуального анализа данных представляет собой механизм, создающий модели интеллектуального анализа данных.



Средства интеллектуального анализа данных в этой надстройке автоматически анализируют распределение и тип данных и рекомендуют лучший способ обработки данных для получения допустимых результатов.

ОБЗОР СРЕДСТВ АНАЛИЗА ТАБЛИЦ ДЛЯ EXCEL

<u>Анализ ключевых факторов влияния</u>	Определяет столбцы данных с наибольшим влиянием на выбранное значение или столбец значений.
<u>Поиск категорий</u>	Определяет строки с похожими свойствами.
<u>Заполнение по примеру</u>	Поиск отсутствующих значений данных в выбранном столбце и предложение новых значений на основе закономерностей в данных.
<u>Прогноз</u>	Прогнозирует будущие значения с учетом ряда значений.
<u>Выделение исключений</u>	Поиск значений в столбце данных, не соответствующих шаблонам, обнаруженным в данных.
<u>Анализ сценария: поиск решения</u>	Указывает целевое значение и определяет базовые факторы, подлежащие изменению для соответствия цели на основе анализа шаблонов данных.
<u>Анализ сценария: гипотетические ситуации</u>	Использует значения для определения результата изменения на основе анализа закономерностей в данных

АНАЛИЗ КЛЮЧЕВЫХ ФАКТОРОВ ВЛИЯНИЯ

	A	B	C	D	E	F
1	Отчет по ключевым факторам влияния для "Occupation"					
2						
3	Ключевые факторы влияния и их воздействие на значения "Occupation"					
4	Отфильтруйте по "Столбец" или "Подходит", чтобы увидеть, как разные столбцы влияют на "Occupation"					
5	Столбец	Значение	Подходит	Относительное влияние		
6	Income	39050 - 71062	Skilled Manual			
7	Region	North America	Skilled Manual			
8	Commute Distance	5-10 Miles	Skilled Manual			
9	Cars	2	Skilled Manual			
10	Age	<37	Skilled Manual			
11	Children	1	Skilled Manual			
12	Education	High School	Skilled Manual			
13	Income	<39050	Clerical			
14	Region	Europe	Clerical			
15	Commute Distance	0-1 Miles	Clerical			
16	Cars	0	Clerical			
17	Education	Partial College	Clerical			
18	Education	Partial High School	Clerical			
19	Age	>= 65	Clerical			
20	Commute Distance	10+ Miles	Professional			
24						
26	Children	5	Professional			
27	Age	46 - 55	Professional			
28	Cars	3	Professional			
29	Children	4	Professional			
30	Income	<39050	Manual			
31	Region	Europe	Manual			
32	Education	Partial High School	Manual			
33	Commute Distance	0-1 Miles	Manual			
34	Education	High School	Manual			
35	Age	<37	Manual			
36	Marital Status	Single	Manual			
37	Children	0	Manual			
38	Cars	1	Manual			
39	Income	97111 - 127371	Management			
40	Age	>=65	Management			
41	Education	Bachelors	Management			
42	Age	55 - 65	Management			
43	Income	>=127371	Management			

При создании отчета, средство выполняет три действия:

1. создает структуру интеллектуального анализа данных, хранящую ключевые сведения о данных;
2. создает модель интеллектуального анализа данных с помощью упрощенного алгоритма Байеса Майкрософт;
3. запускает прогнозирующий запрос для каждой заданной пары атрибутов, чтобы определить факторы, наиболее отличающие эти два целевых атрибута.

СРАВНЕНИЕ ФАКТОРОВ, ВЕДУЩИХ К ЗНАЧЕНИЯМ "SKILLED MANUAL" И "MANAGEMENT"

83	Сравнение факторов, ведущих к значениям "Skilled Manual" и "Management"			
84	Отфильтруйте по "Столбец", чтобы увидеть, как разные значения подходят "Skilled Manual" или "Management"			
85	Столбец	Значение	Подходит Skilled Manual	Подходит Management
86	Income	97111 - 127371		
87	Age	>= 65		
88	Age	55 - 65		
89	Education	Partial College		
90	Income	< 39050		
91	Age	< 37		
92	Income	39050 - 71062		
93	Income	>= 127371		
94	Education	Bachelors		
95	Cars	4		
96	Commute Distance	10+ Miles		
97	Cars	3		
98	Children	0		
99	Education	High School		
100	Cars	0		
101	Education	Graduate Degree		
102	Education	Partial High School		
103	Income	71062 - 97111		
104	Children	4		
105	Children	5		
106	Region	Pacific		
107	Children	1		
108	Age	46 - 55		
109				
110				
111				
112				

Интеллектуальный анализ данных SQL Server - Сравнение, основанное на

Сравнение, основанное на ключевых факторах влияния

Microsoft SQL Server 2008

Отчеты о сравнении

Выберите значения для создания отчетов, показывающих, как ключевые факторы различают эти значения. Можно продолжить создание отчетов для различных пар значений или закрыть это диалоговое окно, чтобы завершить анализ.

Анализируемый столбец: Occupation

Сравнить значение 1: Skilled Manual

со значением 2: Management

Добавление отчета Закрыть

Introduction Table Analysis Tools Sample **Факторы влияния для Occupati** Forecasting Fill From Example Source Data

ПОИСК КАТЕГОРИЙ

1 Обнаружено 5 категорий

2

3 Чтобы переименовать категорию, измените значение "Имя категории" ниже.

4 (Изменения значения "Имя категории" видны в столбце "Категория" исходной таблицы Excel)

Имя категории	Счетчик строк
Категория 1	239
Категория 2	238
Категория 3	245
Категория 4	155
Категория 5	123

11

12

13 Характеристики категории

14 Отфильтруйте таблицу по "Категория", чтобы увидеть характеристики разных категорий

Категория	Столбец	Значение	Относительная важность
Категория 1	Cars	2	
Категория 1	Income	Низкое:39050 - 71062	
Категория 1	Region	North America	
Категория 1	Commute Distance	5-10 Miles	
Категория 1	Occupation	Skilled Manual	
Категория 1	Education	High School	
Категория 1	Children	4	
Категория 1	Commute Distance	10+ Miles	
Категория 1	Age	Высокое:55 - 65	
Категория 1	Age	Очень высокое:>= 65	
Категория 1	Home Owner	Yes	
Категория 1	Purchased Bike	No	
Категория 1	Marital Status	Married	
Категория 1	Children	2	
Категория 1	Occupation	Management	
Категория 1	Occupation	Professional	
Категория 1	Gender	Male	

88

89

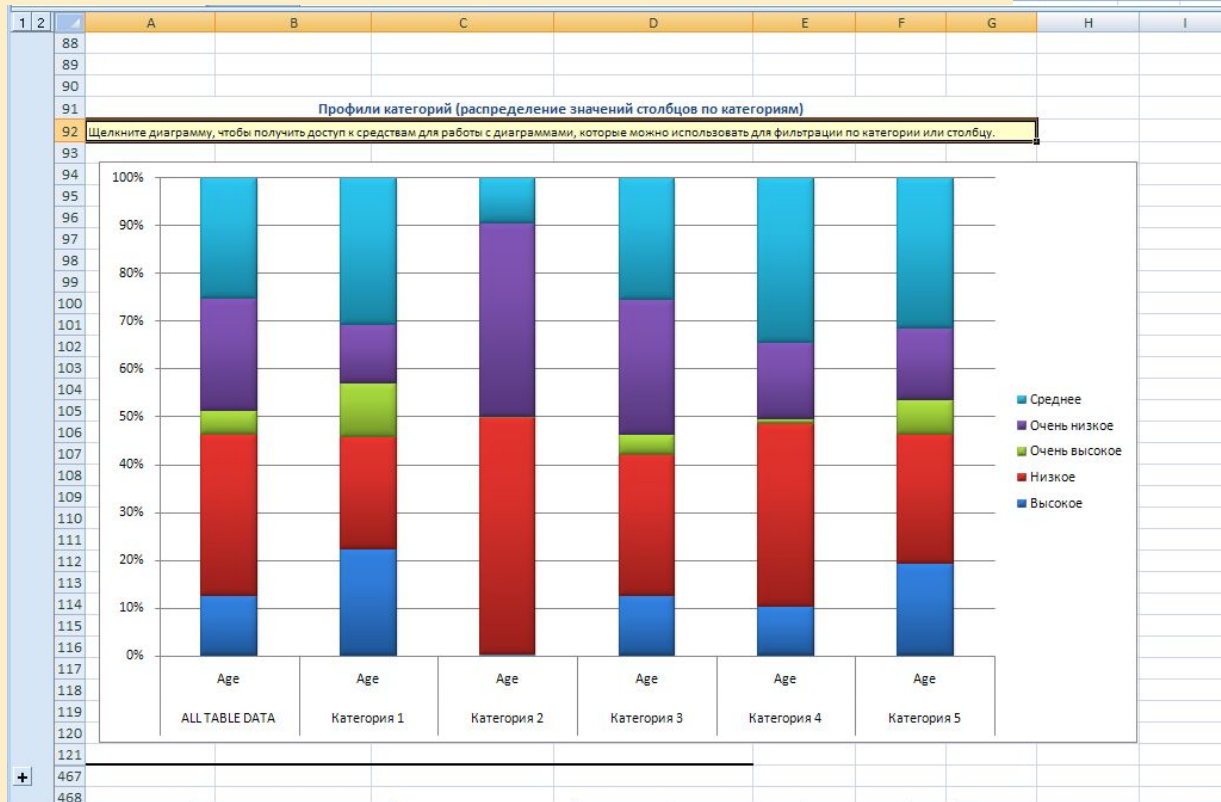
90

Introduction Table Analysis Tools Sample Отчет по категориям Forecasting Fill From Example Source Data Trainin

После завершения работы средства создается отчет со списком найденных категорий вместе с их отличительными характеристиками.

ПОИСК КАТЕГОРИЙ - ДИАГРАММА

• В таблицу данных добавляется новый столбец с предлагаемой категорией



Обнаружение категорий", "Выделение исключений" и "Анализ сценария"

Owner	Cars	Commute Distance	Region	Age	Purchased Bike	Категория
Yes	0	0-1 Miles	Europe	42	No	Категория 2
les			Europe	43	No	Категория 3
les			Europe	60	No	Категория 4
iles			Pacific	41	Yes	Категория 4
les			Europe	36	Yes	Категория 2
les			Europe	50	No	Категория 3
les			Pacific	33	Yes	Категория 5
les			Europe	43	Yes	Категория 2
iles			Pacific	58	No	Категория 3
les			Europe	48	Yes	Категория 3
les			Pacific	54	Yes	Категория 3
les			Pacific	36	No	Категория 4
les			Europe	55	No	Категория 4
les			Europe	35	Yes	Категория 2
les			Pacific	45	Yes	Категория 2
les			Europe	38	Yes	Категория 3
les			Pacific	59	Yes	Категория 3
les			Europe	47	No	Категория 3
les			Europe	35	Yes	Категория 2
iles			Pacific	55	Yes	Категория 3
les			Europe	36	Yes	Категория 2
les			Pacific	35	No	Категория 4
les			Europe	35	Yes	Категория 2
iles			Europe	56	No	Категория 5
les			Europe	34	No	Категория 2
les			Europe	63	No	Категория 3
les			Europe	29	Yes	Категория 3
iles			Pacific	40	No	Категория 4

ЗАПОЛНЕНИЕ ПО ПРИМЕРУ

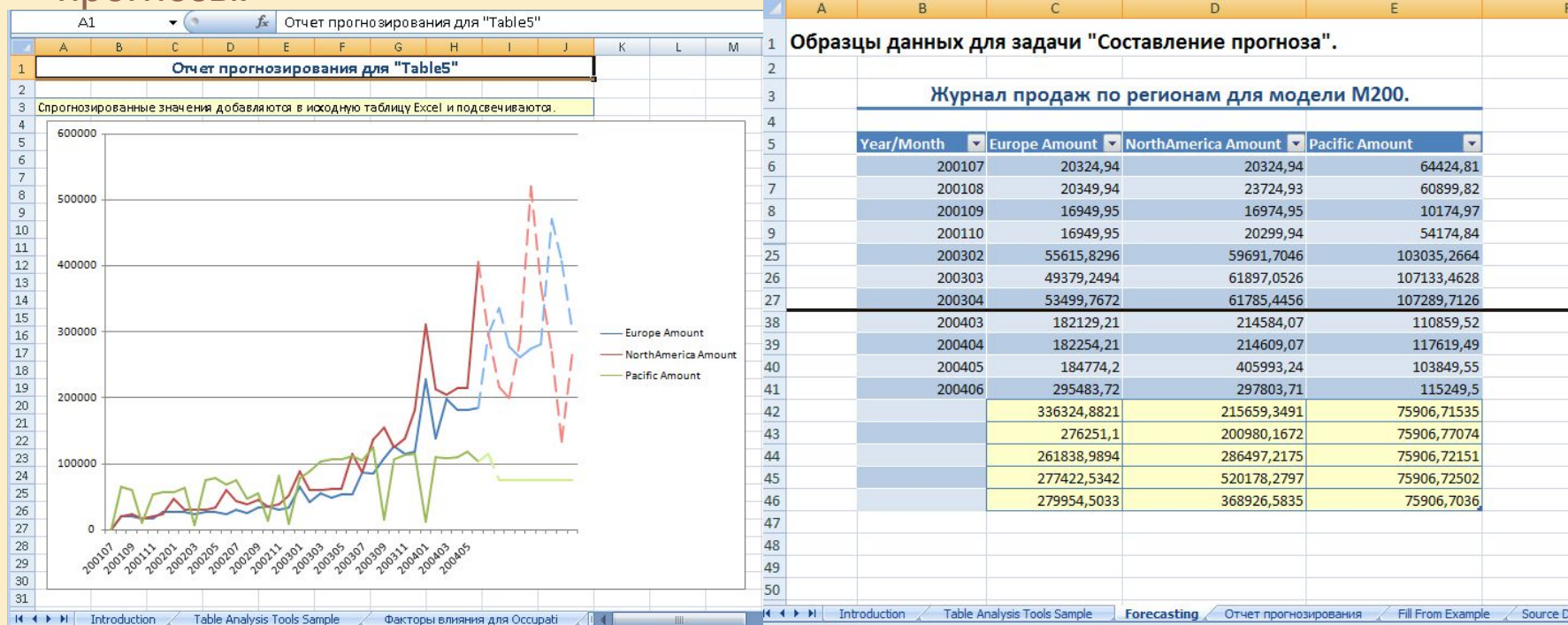
- Средство позволяет быстро создать новые столбцы данных, основанные на закономерностях, найденных в таблице, и образцах новых значений, предоставленных пользователем.

	A	B	C	D	E	F	G
1	Отчет по шаблону для "High Value Customer"						
2							
3	Ключевые факторы влияния и их воздействие на значения "High Value Customer"						
4	Отфильтруйте по "Столбец" или "Подходит", чтобы увидеть, как разные столбцы влияют на "High Value Customer"						
5	Столбец	Значение	Подходит	Относительное влияние			
6	Commute Distance	2-5 Miles	Yes				
7	Children	5	Yes				
8	Region	Europe	Yes				
9	Home Owner	No	Yes				
10	Education	Partial College	Yes				
11	Children	3	Yes				
12	Cars	2	Yes				
13	Education	High School	Yes				
14	Gender	Male	Yes				
15	Occupation	Clerical	Yes				
16	Commute Distance	0-1 Miles	Yes				
17	Occupation	Management	Yes				
18	Region	Pacific	No				
19	Commute Distance	5-10 Miles	No				
20	Gender	Female	No				
21	Education	Partial High School	No				
22	Education	Bachelors	No				
23	Commute Distance	1-2 Miles	No				
24	Occupation	Professional	No				
25	Children	0	No				
26	Children	2	No				
27	Cars	0	No				
28	Home Owner	Yes	No				
29							
30							

	K	L	M	N
ance	Region	Age	High Value Customer	High Value Customer_Extended
	Europe	42	Yes	Yes
	Europe	43	Yes	Yes
	Europe	60	Yes	Yes
	Pacific	41	No	No
	Europe	36	Yes	Yes
	Europe	50	No	No
	Pacific	33	No	No
	Europe	43	Yes	Yes
	Pacific	58	No	No
	Europe	48	Yes	Yes
	Pacific	54		No
	Pacific	36		No
	Europe	55		Yes
	Europe	35		Yes
	Pacific	45		Yes
	Europe	38		Yes
	Pacific	59		Yes
	Europe	47		No
	Europe	35		Yes
	Pacific	55		No
	Europe	36		No
	Pacific	35		No
	Europe	35		Yes
	Europe	56		Yes
	Europe	34		Yes

ПРОГНОЗ

- После завершения мастера новые прогнозы добавятся в конец таблицы источника данных,
- Новые значения рядов времени не добавлены; это позволяет сначала предварительно просмотреть прогнозы.



ВЫДЕЛЕНИЕ ИСКЛЮЧЕНИЙ

• На сводной диаграмме показано число ячеек в каждом столбце, значения в которых превышают порог исключений.

• Средство в исходной таблице выделяет подсветкой ячейки с подозрительными значениями. Темная подсветка означает, что строка требует внимания. Светлая подсветка означает, что значение в этой конкретной ячейке рассматривается как подозрительное.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
	Ячейки выбросов выделены подсветкой в исходной таблице.																			
	Порог исключений (больше или меньше исключений)	75																		
	Столбец	Выбросы																		
	Marital Status	0																		
	Gender	0																		
	Income	0																		
	Children	3																		
	Education	4																		
	Occupation	0																		
	Home Owner	0																		
	Cars	4																		
	Commute Distance	1																		
	Region	1																		
	Age	2																		
	Purchased Bike	0																		
	Итого	15																		

Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	Purchased Bike
d	Male	130000	4	Partial College	Professional	No	4	5-10 Miles	Europe	61	Yes
d	Female	40000	1	Bachelors	Skilled Manual	Yes	0	0-1 Miles	Europe	43	Yes
d	Male	60000	2	Bachelors	Professional	Yes	1	2-5 Miles	Pacific	28	Yes
	Female	10000	1	High School	Manual	No	1	1-2 Miles	Europe	45	No
	Female	10000	2	High School	Manual	Yes	0	0-1 Miles	Europe	35	No
d	Male	40000	2	Bachelors	Management	Yes	1	0-1 Miles	Pacific	52	Yes
	Male	60000	4	Bachelors	Professional	Yes	3	10+ Miles	Pacific	41	No
d	Female	30000	1	Bachelors	Clerical	Yes	0	0-1 Miles	Europe	37	Yes
	Male	30000	2	Partial College	Clerical	Yes	2	5-10 Miles	Pacific	68	No
d	Female	40000	0	Graduate Degree	Clerical	Yes	0	0-1 Miles	Europe	37	Yes
	Male	30000	0	High School	Manual	Yes	1	2-5 Miles	Europe	33	Yes
	Female	20000	4	High School	Manual	Yes	1	0-1 Miles	Europe	43	Yes
d	Female	10000	0	Partial High School	Manual	No	2	0-1 Miles	Europe	30	No
d	Male	120000	0	Partial High School	Professional	Yes	4	10+ Miles	Pacific	36	Yes
	Female	10000	0	Partial High School	Manual	No	2	0-1 Miles	Europe	35	No
d	Female	130000	3	High School	Professional	Yes	4	0-1 Miles	Europe	52	No
	Female	20000	0	Partial College	Manual	No	1	2-5 Miles	Europe	36	Yes
d	Female	20000	3	High School	Skilled Manual	No	2	1-2 Miles	Pacific	62	No
	Female	130000	4	High School	Management	Yes	4	0-1 Miles	Pacific	31	No
	Female	20000	0	Partial High School	Manual	No	2	1-2 Miles	Europe	26	No
d	Male	80000	0	Bachelors	Professional	Yes	2	10+ Miles	Pacific	29	Yes
	Male	80000	2	High School	Skilled Manual	No	2	1-2 Miles	Pacific	50	Yes
	Male	40000	2	Bachelors	Management	Yes	2	5-10 Miles	Pacific	63	Yes
d	Female	30000	4	Graduate Degree	Clerical	Yes	0	0-1 Miles	Europe	45	Yes
	Female	10000	4	Partial High School	Manual	Yes	2	0-1 Miles	Europe	40	No
d	Male	30000	0	Bachelors	Clerical	Yes	0	0-1 Miles	Europe	47	Yes
	Male	20000	0	High School	Manual	No	1	2-5 Miles	Europe	29	No
	Male	40000	2	Bachelors	Management	No	1	5-10 Miles	Pacific	52	Yes
	Male	10000	0	Partial College	Manual	Yes	1	1-2 Miles	Pacific	26	Yes
	Male	130000	3	Partial College	Professional	No	3	0-1 Miles	Europe	51	Yes
d	Male	80000	5	Bachelors	Professional	Yes	4	1-2 Miles	Pacific	40	No
	Male	30000	0	Partial College	Clerical	No	1	2-5 Miles	Europe	29	No
d	Male	20000	1	High School	Manual	No	1	1-2 Miles	Europe	40	Yes
	Female	30000	0	Partial College	Clerical	No	1	0-1 Miles	Europe	26	Yes

АНАЛИЗ СЦЕНАРИЯ ПОИСК РЕШЕНИЯ

Сценарий Поиск решения представляет собой дополнение к средству сценария **Анализ**

гипотетических вариантов и указывает на влияющие факторы, которые должны быть изменены

При создании сценария поиска решения выполняются следующие действия.

1. Создает структуру интеллектуального анализа данных, в которой хранятся ключевые сведения о содержащихся в таблице данных.
2. На основе существующих данных создает модель интеллектуального анализа с логистической регрессией.
3. Создает прогнозирующий запрос для каждого из указанных значений.

	C	D	E	H	I	L	M	N	O
3	Gender	Income	Children	Home Owner	Cars	Age	Purchased Bike	Цель: Cars=2	Рекомендованный столбец: Income
4	Female	40000	1	Yes	0	42	No	⊗	6365
5	Male	30000	3	Yes	1	43	No	⊙	10000
6	Male	80000	5	No	2	60	No	⊙	80000
7	Male	70000	0	Yes	1	41	Yes	⊗	6365
8	Male	30000	0	No	0	36	Yes	⊗	6365
9	Female	10000	2	Yes	0	50	No	⊗	6365
10	Male	160000	2	Yes	4	33	Yes	⊙	10000
11	Male	40000	1	Yes	0	43	Yes	⊗	6365
12	Male	20000	2	Yes	2	58	No	⊙	20000
13	Male	20000	2	Yes	1	48	Yes	⊗	6365
14	Female	30000	3	No	2	54	Yes	⊙	30000
15	Female	90000	0	No	4	36	No	⊗	6365
16	Male	170000	5	Yes	4	55	No	⊙	10000
17	Male	40000	2	Yes	1	35	Yes	⊗	6365
18	Male	60000	1	No	1	45	Yes	⊗	6365
19	Female	10000	2	Yes	1	38	Yes	⊗	6365
20	Male	30000	3	No	2	59	Yes	⊙	30000
21	Female	30000	1	Yes	0	47	No	⊗	6365
22	Male	40000	2	Yes	1	35	Yes	⊗	6365
23	Male	20000	2	Yes	2	55	Yes	⊙	20000
24	Female	40000	0	Yes	0	36	Yes	⊗	6365
25	Female	80000	0	Yes	4	35	No	⊗	6365
26	Male	40000	2	Yes	0	35	Yes	⊗	6365
27	Female	80000	5	No	3	56	No	⊙	10000
28	Male	40000	2	No	1	34	No	⊗	6365

Introduction Table Analysis Tools Sample Forecasting Отчет прогнозирования Fill From Example Source Data Training Data Testing Data

АНАЛИЗ СЦЕНАРИЯ: ГИПОТЕТИЧЕСКИЕ СИТУАЦИИ

Сценарий анализирует закономерности существующих данных, а затем позволяет оценить влияние изменений в одном столбце на значение другого столбца.

	L	M	N	O	P	Q
1	ние исключений" и "Анализ сценария".					
2						
3	Age	Purchased Bike	Цель: ChildGen5	Рекомендованный столбец: Income	Новое значение: Income	Доверие
4	42	No	✗	6385	48002	
5	43	No	✗	6385	35118	
6	60	No	✓	80000	69355	
7	41	Yes	✗	115780	66781	
8	36	Yes	✗	155416	32649	
9	50	No	✗	6385	19070	
10	33	Yes	✗	24361	112724	
11	43	Yes	✗	6385	51453	
12	58	No	✗	6385	21197	
13	48	Yes	✗	6385	25804	
14	54	Yes	✗	6385	41389	
15	36	No	✗	155416	103886	
16	55	No	✓	170000	114062	
17	35	Yes	✗	6385	40738	
18	45	Yes	✗	6385	56578	
19	38	Yes	✗	6385	13707	
20	59	Yes	✗	6385	30401	
21	47	No	✗	6385	25153	
22	35	Yes	✗	64272	39362	
23	55	Yes	✗	6385	24027	
24	36	Yes	✗	33294	39604	
25	35	No	✗	155416	101737	
26	35	Yes	✗	34736	36679	
27	56	No	✓	80000	93226	
28	34	No	✗	155416	38408	
29	63	No	✗	6385	20001	
30	29	Yes	✗	155416	43118	
31	40	No	✗	155416	65730	
32	44	No	✓	70000	53115	
33	32	Yes	✗	155416	22582	
34	63	No	✗	6385	15642	
35	26	Yes	✗	155416	29640	
36	31	No	✗	155416	16397	
37	50	Yes	✗	6385	41271	
38	62	Yes	✓	90000	70926	
39	41	No	✓	10000	16866	
40	50	Yes	✗	6385	25048	
41	30	No	✗	155416	39264	

При создании сценария средство выполняет задачи:

1. Создает структуру интеллектуального анализа данных, в которой хранятся ключевые сведения о содержащихся в таблице данных.
2. На основе существующих данных создает модель интеллектуального анализа с логистической регрессией.
3. Создает прогнозирующий запрос для каждого из указанных значений.

СПАСИБО