

Лабораторная работа №12. Анализ данных в среде Deductor.

1. В данной лабораторной работе студенту предлагается ознакомиться с возможностями, предоставляемыми средой анализа данных Deductor (<https://basegroup.ru/deductor>).
2. Необходимо установить предлагаемую среду, провести манипуляции над данными согласно заданию (начинается со следующей страницы) и оформить письменный отчет.
3. В качестве набора данных для анализа можно взять предлагаемый в самом задании набор данных либо любой другой по выбору студента, **позволяющий полностью выполнить описанные в задании действия.**
4. Отчет оформляется в свободной форме и должен содержать:
 - a) ФИО и группу студента;
 - b) Название дисциплины и лабораторной работы;
 - c) Краткую формулировку задания;
 - d) Краткую характеристику выбранного набора данных (где взят, какие объекты описывает и сколько их, какими признаками эти объекты обладают);
 - e) Пошаговое описание действий, **сопровожаемое скриншотами.**
5. Отчеты, в которых отсутствуют один или несколько элементов из списка выше, не засчитываются.
6. Данная лабораторная не предполагает очной защиты.

Задание 1

Используя мастер импорта, загрузить в среду Deductor Studio данные из текстового файла «Пример call-центр.txt».

Задание 2

Используя мастер обработки, провести очистку данных: выявление дубликатов и противоречий. В качестве входных атрибутов использовать атрибуты «Занятость», «Частота пользования услугами», «Эмоциональность», «Удовлетворенность». В качестве выходных атрибутов использовать атрибут «Действующий клиент».

Мастер обработки - Дубликаты и противоречия (1 из 4)

Выявление дубликатов и противоречий

Выявление дубликатов и противоречий

Имя столбца: COL13

Метка столбца: Действующий клиент

Тип данных: ab Строковый

Вид данных: Дискретный

Назначение: Выходное

Действующий клиент

< Назад Далее > Отмена

В результате будут выявлены дубликаты и противоречия:

Дубликаты		Противоречия		Выходные поля				Выходные поля				Информ	
Признак	Группа	Признак	Группа	Занятость	Частота пользования услугами	Эмоциональность	Удовлетворенность	Действующий клиент	Фамилия	Имя	Отчество	Пол	Воз
✓	5	✓		работает	средняя	средняя	высокая	Да	Крылов	Петр	Юрьевич	м	
✓	2	✓		без работы	средняя	высокая	высокая	Да	Жихарева	Ксения	Леонидовна	ж	
✓	3	✓		на пенсии	средняя	высокая	низкая	Нет	Даматов	Сергей	Михайлович	м	
✓	5	✓		работает	средняя	высокая	высокая	Да	Артурова	Елена	Анатовна	ж	
✓	4	✓		работает	средняя	низкая	низкая	Да	Шаркин	Станислав	Артурович	м	
✓	4	✓		работает	средняя	низкая	низкая	Да	Шаркина	Юлия	Петровна	ж	
✓	2	✓		без работы	средняя	высокая	высокая	Да	Шалагин	Денис	Алексеевич	м	
✓	3	✓		на пенсии	средняя	высокая	низкая	Нет	Дружина	Мария	Анатовна	ж	
✓	1	✓		без работы	низкая	низкая	низкая	Нет	Базанов	Артём	Валерьевич	м	
✓		✓	1	на пенсии	низкая	высокая	высокая	Нет	Киселева	Диана	Анатовна	ж	
✓	1	✓		без работы	низкая	низкая	низкая	Нет	Халева	Екатерина	Артуровна	ж	
✓		✓	1	на пенсии	низкая	высокая	высокая	Да	Гаврилов	Ян	Александрович	м	

Задание 3

Используя мастер обработки, провести квантование данных. Возраст и дату первого заказа необходимо разделить на три интервала:

Мастер обработки - Квантование (1 из 4)

Квантование
Настройка параметров квантования

☐ Фамилия
☐ Имя
☐ Отчество
☐ Пол
☒ Возраст
☐ Занятость
☒ Дата первого заказа
☐ Город
☐ Частота пользования услугами
☐ Эмоциональность
☐ Удовлетворенность
☐ Способы оплаты
☐ Действующий клиент

Имя столбца: COL5
Тип данных: Вещественный
Назначение: ☒ Используемое
Способ: По квантилям
Интервалов: 3
Значение: Автоматическая метка
Вид данных: ... Дискретный
Минимум: 19
Максимум: 80
Стандартное откл.: 17,9937517500052

< Назад Далее > Отмена

Мастер обработки - Квантование (1 из 4)

Квантование
Настройка параметров квантования

☐ Фамилия
☐ Имя
☐ Отчество
☐ Пол
☒ Возраст
☐ Занятость
☒ Дата первого заказа
☐ Город
☐ Частота пользования услугами
☐ Эмоциональность
☐ Удовлетворенность
☐ Способы оплаты
☐ Действующий клиент

Имя столбца: COL7
Тип данных: Дата/Время
Назначение: ☒ Используемое
Способ: По квантилям
Интервалов: 3
Значение: Автоматическая метка
Вид данных: ... Дискретный
Минимум: 01.01.2005
Максимум: 09.08.2012
Стандартное откл.: 793дн. 13:53:08


< Назад Далее > Отмена

Мастер обработки - Квантование (2 из 4)

×

Границы и метки интервалов

Настройка границ и меток интервалов квантования



Столбцы	
Имя	Интервалов
9.0 Возраст	3
7 Дата первого заказа	3

Интервалы		
Nº	Граница	Метка
	19	
0	25	до 25
1	48	от 25 до 48
2	80	от 48

< Назад

Далее >


Отмена

Мастер обработки - Квантование (2 из 4)

×

Границы и метки интервалов

Настройка границ и меток интервалов квантования



Столбцы	
Имя	Интервалов
9.0 Возраст	3
7 Дата первого заказа	3

Интервалы		
Nº	Граница	Метка
	01.01.2005	
0	05.08.2007	до 05.08.2007
1	03.02.2010	от 05.08.2007 до 03.02.2010
2	09.08.2012	от 03.02.2010

< Назад

Далее >

Отмена

Задание 4

Используя мастер обработки, построить дерево решений. Атрибуты «Фамилия», «Имя» и «Отчество» задать в качестве информационных. Атрибут «Действующий клиент» установить в качестве выходного атрибута. Остальные атрибуты установить в качестве входных атрибутов:

Мастер обработки - Дерево решений (1 из 7)

Настройка назначений столбцов
Задайте назначения исходных столбцов данных

Фамилия
Имя
Отчество
Пол
Возраст
Занятость
Дата первого заказа
Город
Частота пользования услугами
Эмоциональность
Удовлетворенность
Способы оплаты
Действующий клиент

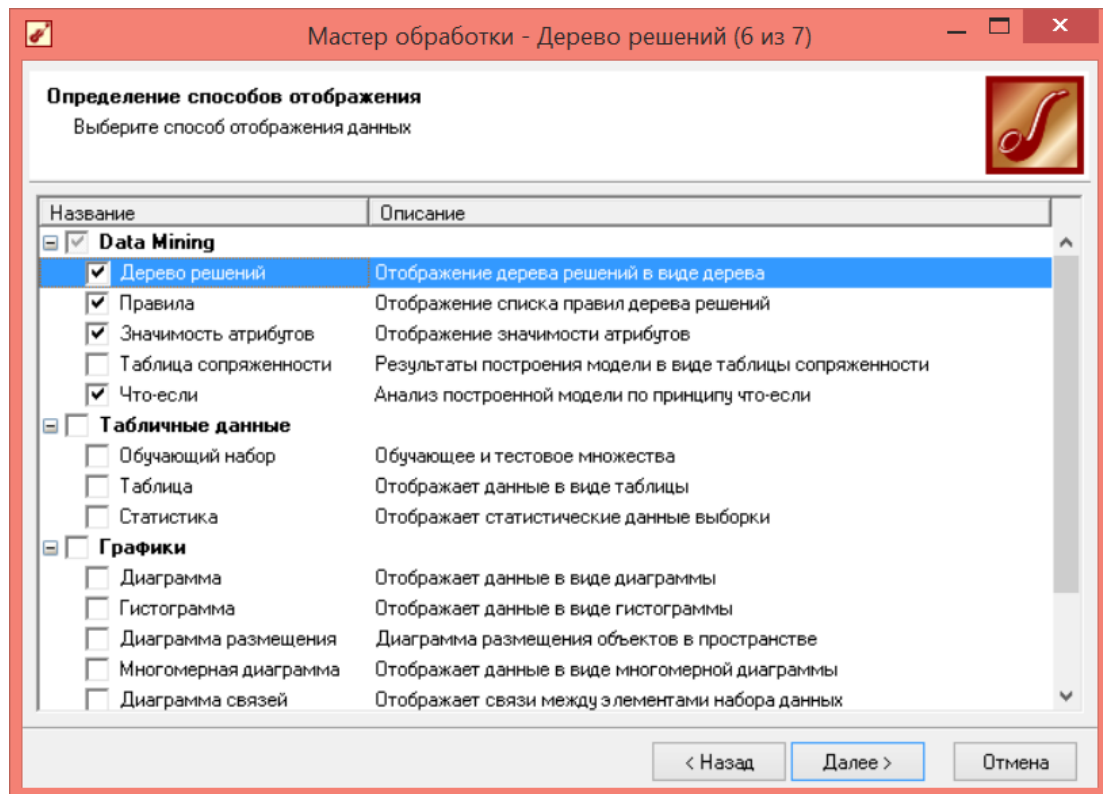
Имя столбца: COL4
Тип данных: Строковый
Назначение: Входное
Вид данных: Дискретный

Уникальные значения
Кол-во уникальных значений: 2
ж
м

Настройка нормализации...

< Назад Далее > Отмена

Отобразить полученные результаты в качестве дерева решений, правил, значимости атрибутов и «что-если» анализа:



Задание 5

Используя «что-если» анализ, спрогнозировать вероятность потери клиента с заданными характеристиками. Характеристики клиента задать самостоятельно.

Дерево решений X Правила X Значимость атрибутов X Таблица сопряженности X Что-если X	
1 из 50	
Поле	Значение
Входные	
ab Пол	м
ab Возраст	от 25 до 48
ab Занятость	работает
ab Дата первого заказа	от 05.08.2007 до 03.02.2010
ab Город	Березники
ab Частота пользования услугами	средняя
ab Эмоциональность	низкая
ab Удовлетворенность	высокая
ab Способы оплаты	перевод
Выходные	
ab Действующий клиент	Да
Расчетные	
12 Действующий клиент Номер правила	8
9.0 Действующий клиент Поддержка, %	14,5833333333333
9.0 Действующий клиент Достоверность, %	100

Вспомогательный материал

Описание предметной области. Call-центр – это система эффективной обратной связи с потребителем товаров и услуг либо поддержки, продвижения различных акций, социальных опросов, голосований.

Call-центры бывают двух типов:

- Аутсорсинговый call-центр;
- Корпоративный call-центр;

Аутсорсинговые call-центры бывают двух видов. При первом варианте схемы организации аутсорсингового call-центра услуги сдаются в аренду с подробным обучением операторов особенностям консультирования клиентов только по необходимым для арендующей компании вопросам. При второй вариант аутсорсингового call-центра организация происходит не на постоянной основе, а однократно или от случая к случаю.

Корпоративный call-центр - это внутреннее штатное подразделение компании, выполняющее функции работы с партнерами и клиентами. Основные затраты при организации собственного call-центра приходятся на аренду помещения, закупку системы оборудования, подбор и обучение персонала.

К основным задачам call-центра относятся:

- Приема поступающей информации
- Обработки информации
- Оперативное реагирование на изменения в системе вызовов
- Создание и изменение базы данных клиентов call-центра
- Обновление программного и аппаратного обеспечения
- Регулярное обучение персонала
- Ведение статистики
- Маршрутизация вызовов
- Взаимодействие с другими отделами компании
- Улучшение обслуживания.

Можно выделить два приоритетных направления работы call-центра: входящая и исходящая связь. К основным задачам входящей относится предоставление клиентам необходимой информации, а также получение сведений о проблемах, которые возникают у клиентов при использовании товарами или услугами компании.

К основным задачам исходящей связи относятся продажа товаров или услуг, информирование клиентов о нововведениях, и проведение опросов.

К возможностям call-центров относятся:

- Регистрация звонков
- Хранение информации о клиенте
- Маршрутизация вызовов.
- Запись разговоров
- Графическое отображение работы каждого оператора или отдела
- Отображение на мониторе оператора информации о поступившем вызове
- Переадресации, создание очереди звонков, включение режима ожидания, автоматическое информирование о времени ожидания ответа
- Распределение звонков внутри группы операторов в зависимости от статуса.

На сегодняшний день основной проблемой большинства call-центров или многих других организаций, главная деятельность которых есть взаимодействие с людьми, являются сложности, во-первых, в удерживании имеющихся клиентов и поддержки их лояльности, во-вторых, трудности в поиске и привлечение новых клиентов. Все это является следствием неэффективного и плохого сервиса обслуживания клиентов. Следовательно, повышать эффективности деятельности call-центров следует через улучшение качества обслуживания клиентов.

Очистка данных: дубликаты и противоречия. Для удобства работы с данными для начала необходимо произвести очистку данных. Предварительная обработка анализируемых данных является необходимым шагом, так как для

исходных или «сырых» данных чаще всего необходима очистка, потому что они не соответствуют определенным критериям качества. Если избежать этапа предварительной обработки, то результаты анализа будут неточными. Потребность в предварительной обработке возникает независимо от используемых алгоритмов и технологий.

Дубликаты - записи в таблице, все входные и выходные поля которых одинаковые.

Противоречия - записи в таблице, у которых все входные поля одинаковые, но отличаются хотя бы по одному выходному полю.

При построении модели классификации необходимо определить входные и выходные поля, зависимости между которыми будут исследоваться. При этом значения входных полей должны полностью определять значения выходных. Такая постановка задачи часто приводит к возникновению противоречий, то есть к наличию записей, с одинаковыми значениями входных полей, но разными значениями в выходных полях. Чаще всего только одна запись в найденных противоречиях является правильной, а присутствующие ошибочные записи искажают результаты, поэтому противоречия лучше полностью исключить из исходной анализируемой выборки. Но не всегда наличие противоречий оказывает негативное влияние на результаты исследования. Противоречия могут быть полезны, если есть необходимость введения неопределенности в данные, более того, противоречия могут отражать особенности анализируемого объекта.

Если в данных встречаются записи с одинаковыми входными и выходными значениями, то говорят о дубликатах. Дубликаты – это избыточные данные, которые чаще всего указывают на ошибку при подготовке исходной выборки. Следовательно, дубликаты могут исказить результаты анализов. С другой стороны, дубликаты, как и противоречия, могут быть полезны при проведении анализа, так как они могут указывать на значимость дублирующейся информации.

Квантование данных. Квантование – это операция, при выполнении которой происходит разбиение диапазона числовых значений на установленное количество интервалов и замена каждого значения на число, связанное с интервалом, к которому оно относится. Данные интервалы включают в себя только нижнюю границу, за исключением последнего интервала, который включает в себя как нижнюю, так и верхнюю границу. Результатом квантования может быть:

- Номер интервала
- Значение нижней или верхней границы интервала
- Среднее значение интервала
- Метка интервала
- Автоматическая метка.

Деревья решений. Метод «Деревья решений» предоставляет возможность решения задач классификации и прогнозирования. Задачи классификации имеют место быть, если зависимая переменная принимает дискретные значения. Если зависимая переменная принимает непрерывные значения, то при помощи дерева решений возможно решение задач прогнозирования. Пример дерева решений:

Условие	Следствие	Поддержка	Достоверность
ЕСЛИ		59	31
Частота пользования услугами = высокая	False	16	16
Частота пользования услугами = низкая	True	19	19
Частота пользования услугами = средняя		24	12
Удовлетворенность = высокая	False	6	6
Удовлетворенность = низкая		8	7
Возраст = до 28	True	4	4
Возраст = от 28 до 55	False	1	1
Возраст = от 55	True	3	3
Удовлетворенность = средняя		10	5
Занятость = без работы	True	2	2
Занятость = домохозяйка	False	2	1
Занятость = на пенсии	True	2	2
Занятость = работает	False	2	2
Занятость = учится	False	2	2

Согласно результатам, представленным в примере дерева решений, можно сделать следующие выводы:

- Если частота пользования услугами высокая, то вероятность покинуть целевую аудиторию отсутствует. Если низкая, то вероятность

покинуть целевую аудиторию высокая. Следовательно, имеет смысл уделять больше внимания клиентам, которые реже всего пользуются услугами, если существует потребность в данных клиентах. Для этого возможно проведение акции, использование коммерческих предложений, различных поощрений.

- Если частота пользования услугами средняя, то все зависит от удовлетворенности клиента: чем удовлетворенность выше, тем вероятность ниже покинуть целевую аудиторию ниже.
- Если удовлетворенность низкая, то вероятность выхода из целевой аудитории зависит от возраста. До 28 лет и после 55 лет вероятность возрастает, в промежутке от 28 до 55 вероятность покинуть целевую аудиторию низкая.
- Также есть зависимость с родом деятельности клиента. Например, для пенсионеров и безработных высока вероятность прекращения пользования данными услугами, поэтому не имеет смысла уделять много внимания данным клиентам и привлекать их к пользованию услугами.

На основе данного дерева решений можно сделать вывод, что самыми привлекательными клиентами являются работающие, учащиеся и домохозяйки.