МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

СОЗДАНИЕ ПРИЛОЖЕНИЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ ПРОДАЖИ АВИАБИЛЕТОВ С ПОМОЩЬЮ ТЕХНОЛОГИИ DATA MINING

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 5 курса 551 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Бисенгалиевой Асель Насивулловны

Научный руководитель
Старший преподаватель _____ М. И. Сафрончик

Заведующий кафедрой
к. ф.-м. н., доцент _____ А. С. Иванов

Саратов 2020

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Повсеместное использование баз данных и накопление огромного объема информации привело к пониманию необходимости его всестороннего анализа с целью получения новых знаний и выработки эффективных стратегий дальнейшего развития. Возникла потребность в программных инструментах, позволяющих удобным для пользователя образом извлекать значимую информацию, на ее основе проводить анализ, вырабатывать стратегию и принимать различного рода решения.

На сегодняшний день именно технологии Data Mining позволяют проводить глубинный анализ и выявлять скрытые закономерности в огромном объеме накопленной информации.

Целью работы является создание решения по практическому применению методов классификации Data mining для прогнозирования.

В ходе работы необходимо решить следующие задачи:

- проанализировать предметную область;
- создать представление источника данных;
- создать модель интеллектуального анализа данных;
- обучить модель с помощью алгоритмов дерева принятия решений, кластеризации и упрощенного алгоритма Байеса;
- исследовать и проверить модель (тестирование);
- создать прогноз на основе построенной модели интеллектуального анализа данных.

Программные средства использованные в данной работе: Microsoft SQL Server 2014 и Microsoft Visual Studio 2017.

1 Data Mining

Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей.

Термин Data Mining часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, "извлечение зерен знаний из гор данных раскопка знаний в базах данных, информационная проходка данных, "промывание"данных. Понятие "обнаружение знаний в базах данных "(Knowledge Discovery in Databases, KDD) можно считать синонимом Data Mining [1].

Процесс Data Mining может быть успешным и неуспешным. Использование Data Mining не является гарантией получения исключительно достоверных знаний и принятия на основе этих знаний абсолютно верных решений.

Построенная модель может обладать рядом погрешностей: недостоверные исходные допущения при построении модели; ограниченные возможности при сборе необходимых данных; неуверенность и страхи пользователя системы, и, в силу этого, слабое их применение; неоправданно высокая стоимость [2].

2 Алгоритмы интеллектуального анализа данных

2.1 Алгоритм дерева принятия решений

Согласно наиболее общему определению, дерево принятия решений – это средство поддержки принятия решений при прогнозировании, широко применяющееся в статистике и анализе данных [3].

Алгоритм Microsoft дерева принятия решений — это алгоритм классификации и регрессии, предоставляемый службой Microsoft SQL Server Analysis Services для использования в прогнозном моделировании как дискретных, так и непрерывных атрибутов.

Алгоритм дерева принятия решений (Microsoft) строит модель интеллектуального анализа данных путем создания ряда разбиений в дереве. Эти разбиения представлены как узлы. Алгоритм добавляет узел к модели каждый раз, когда выясняется, что входной столбец имеет значительную корреляцию с прогнозируемым столбцом. Способ, которым алгоритм определяет разбиение, отличается в зависимости от того, прогнозирует ли он непрерывный столбец или дискретный столбец [4].

2.2 Алгоритм кластеризации

Алгоритм Microsoft кластеризации последовательностей — это алгоритм анализа последовательностей, предоставляемый службой Microsoft SQL Server Analysis Services.

Алгоритм кластеризации использует итеративный метод группировки записей набора данных в кластеры, обладающие сходными характеристиками. Используя разбиение на кластеры можно выявить в исследуемом массиве данных такие связи, которые невозможно обнаружить простым просмотром этих данных. Кроме того, с помощью алгоритмов кластеризации можно осуществлять прогнозирование. К примеру, объединить в группу людей, которые живут в одном районе, водят одну марку машин, имеют сходные предпочтения в пище и покупают один тип продукции. Такое объединение и есть кластер. Другой кластер может включать в себя людей, посещающих один ресторан, имеющих один уровень дохода и ездящих дважды в год в отпуск в другие страны.

Оценивая распределение данных в этих кластерах, можно лучше понять взаимосвязи различных характеристик исследуемых объектов, а также как эти взаимосвязи влияют на значение прогнозируемого атрибута [5].

2.3 Упрощенный алгоритм Байеса

Упрощенный алгоритм Байеса от Microsof — это алгоритм классификации, основанный на теореме Байеса и представляемый для использования в прогнозном моделировании в службах Microsoft SQL Server Analysis Services. Слово «упрощенный» в его названии указывает на то, что алгоритм использует методы Байеса, но не учитывает возможные зависимости [3].

Обучение байесовских сетей стало одним из актуальных направлением вычислительной математики и до сих пор является предметов активных исследований [6].

Однако, до сих пор определение структуры байесовской сети в общем виде является сложной задачей как с теоретической, так и с вычислительной точки зрения. Подход в общем виде обладает следующими недостатками [7]:

- Вычислительная сложность.
- При попытке учесть большое количество зависимостей между переменными, оценки условных вероятностей приобретают большую дисперсию, так как их совместное появление в данных является маловероятным событием. Таким образом, оценки параметров могут стать недостоверными, что в итоге может приводить к ухудшению качества классификации даже по сравнению с «наивным» алгоритмом Байеса.
- Из-за большого количества параметров, модель получается слишком ориентированной на обучающие данные. Это приводит к очень хорошим результатам классификации на обучающих данных и неудовлетворительным результатам на тестовых данных. Т.е. модель описывает не общие закономерности в структуре данных, а скорее набор частных случаев в обучающей выборке.

3 Язык DMX

Расширения интеллектуального анализа данных (DMX) — это язык, который можно использовать для создания моделей интеллектуального анализа данных и работы с ними в службах Microsoft SQL Server Analysis Services. Расширения интеллектуального анализа данных могут использоваться для создания структуры новых моделей интеллектуального анализа данных, обучения этих моделей, а также для просматривания, управления и прогнозирования по этим моделям. Расширения интеллектуального анализа данных состоят из инструкций языка определения данных (DDL), инструкций языка обработки данных (DML), а также функций и операторов.

Во всех прогнозирующих запросах используется DMX — язык расширений интеллектуального анализа данных. Синтаксис расширения интеллектуального анализа данных похож на синтаксис T-SQL, но служит для написания запросов к объектам интеллектуального анализа данных [8].

4 Реализация приложения

4.1 Создание представления источника данных

Проектируемый источник данных создается на основе уже существующей базы данных оперативного доступа в среде Microsoft SQL Server 2014. В качестве предметной области были взяты авиаперевозки по России [9].

Основной задачей на данном этапе является определение бизнес процессов, которые требуют углубленного анализа, определение ключевых показателей, по которым будут анализироваться эти бизнес процессы, а так же избавление от лишних связей между сущностями базы данных, которые в первоначальном своем варианте были высоконормализированными.

В приложении нам потребуется выполнить сценарий рассылки почты, в которой используется машинное обучение для анализа и прогнозирования поведения заказчика. Созданный сценарий покажет, как работают основные алгоритмы интеллектуального анализа данных: деревья принятий решений, кластеризации и упрощенный алгоритм Байеса. Также результаты будут проанализированы с помощью средств просмотра модели и созданы прогнозы и диаграммы точности с помощью средств интеллектуального анализа данных, входящих в службы Microsoft SQL Server Analysis Services [10].

Для проекта понадобятся две таблицы `tickets` и `tick`.

Создав проект Analysis Servises в среде Visual Studio 2017, необходимо сначала связать его с `источником данных`. В данном проекте это база `Avia`, которая содержится на сервере `1-PC`.

Далее необходимо создать представление источников данных. Оно формируется на основе хранилища и позволяет выбрать только интересующие нас компоненты хранилища. Далее можно приступать к созданию модели интеллектуального анализа.

4.2 Создание модели интеллектуального анализа данных

На первом шаге создания сценария Targeted Mailing потребуется использование мастера интеллектуального анализа данных среды SQL Server Data Tools (SSDT), на основе существующей реляционной базы данных. Для структуры выберем алгоритм дерева принятия решений.

Для формирования данной структуры следует выбрать таблицы и представления, а затем указать столбцы для обучения и проверочные столбцы. Выбираем входную таблицу tickets и столбцы, а также определим для них типы содержимого и данных. В завершении работы мастера назовем структуру Targeted Mailing, а дерево принятия решений TM_Decision_Tree.

Выберем модель кластеризации, откроем вкладку модель кластеризации, для переменной заливки задаем Buyer, а состояние =1, и переименуем кластер с наибольшей плотностью на "покупатели билета велико". Чем темнее кластер, тем он содержит больший процент числа покупателей. Аналогично поступим для кластера с наименьшей плотностью. Перейдем на вкладку Профили кластера и перетащим в левую часть кластеры с наибольшей и наименьшей плотностью.

Далее перейдем на вкладку Сравнение кластеров, с помощью нее можно исследовать отличие характеристик друг от друга. Отсюда можно сделать вывод, для покупателей с двумя детьми в приоритете заработная плата, а уже после возраст.

Последняя рассматриваемая модель - модель упрощенного алгоритма Байеса. Модель предоставляет несколько методов для отображения взаимодействия между покупателем и входными атрибутами, возраст и доход не были учтены из-за типа. Перейдем на вкладку Сеть зависимостей и заметим, что покупка билета в основном зависит от количества детей. Рассмотрим вкладки Профили атрибутов и Характеристики атрибута.

4.3 Тестирование моделей

Теперь создадим проверку модели с фильтром, с помощью ранее созданного дерева решения создадим деревья решений с фильтрами для мужчин и для женщин. На вкладке `Диаграмма точности прогнозов` можно заметить, что все три модели дерева принятия решений демонстрируют заметную точность по сравнению с моделью случайного выбора, а также превосходят по точности модели на основе алгоритма кластеризации и упрощенного алгоритма Байеса.

4.4 Создание прогноза на основе модели интеллектуального анализа данных

Главный шаг при созданнии прогнозируещего запроса - выбор модели интеллектуального анализа данных и входной таблицы. Для модели выберем раннее созданное дерево решений `TM_Decision_Tree`, а для входной таблицы возьмем более похожую на таблицу вариантов `tick`.

Создадим именованное вычисление для таблицы `tick` по дате рождения и запишем в поле выражения для подсчета возраста.

Перейдем в конструкор, чтобы спроектировать прогнозирующий запрос. Для функции прогнозирования выберем `PredictProbability`, псевдоним: вероятность результата, а критерий или аргумент возьмем `Buyer`. После нажатия на кнопку результат, можно просмотреть результирующий набор данных.

Чтобы сохранить данные результирующей таблицы, щелкнем на "сохранить результат запроса в диалоговом окне выберем имя для новой таблицы с результатом, назовем ее `results`, источник данных и представление источника данных выберем из существующих.

ЗАКЛЮЧЕНИЕ

В ходе работы были решены следующие задачи:

- проанализирована предметная область;
- создано представление источника данных;
- созданы модели интеллектуального анализа данных;
- обучены модели дерева принятия решений, кластеризации и упрощенного алгоритма Байеса;
- исследованы и проверены модели (тестирование);
- создан прогноз на основе построенной модели интеллектуального анализа данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Data Mining [Электронный ресурс].— URL: https://www.intuit.ru/studies/courses/6/6/lecture/158 (Дата обращения 15.05.2020). Загл. с экр. Яз. рус.

2 Процесс Data Mining. Построение и использование модели [Электронный ресурс].— URL: https://www.intuit.ru/studies/educational_groups/1516/courses/6/lecture/196?page=6 (Дата обращения 15.05.2020). Загл. с экр. Яз. рус.

3 Деревья решений и алгоритмы их построения [Электронный ресурс].— URL: http://datareview.info/article/derevya-resheniy-i-algoritmyi-ih-postroeniya/ (Дата обращения 15.05.2020). Загл. с экр. Яз. рус.

4 Алгоритм дерева принятия решений (Майкрософт) [Электронный ресурс].— URL: https://docs.microsoft.com/ru-ru/sql/analysis-services/data-mining/microsoft-decision-trees-algorithm?view=sql-server-2014 (Дата обращения 15.05.2020). Загл. с экр. Яз. рус.

5 Алгоритмы Data Mining [Электронный ресурс].— URL: http://www.businessdataanalytics.ru/DataMiningSQLServer2005-1.htm (Дата обращения 16.05.2020). Загл. с экр. Яз. рус.

6 Модифицированный древовидный алгоритм Байеса [Электронный ресурс].— URL: http://www.businessdataanalytics.ru/DataMiningSQLServer2005-2.htm (Дата обращения 16.05.2020). Загл. с экр. Яз. рус.

7 Модифицированный древовидный алгоритм Байеса для решения задач классификации [Электронный ресурс].— URL: http://businessanalytics.ru/AugmentedNaiveBayes.htm (Дата обращения 16.05.2019). Загл. с экр. Яз. англ.

8 Data Mining Extensions (DMX) Reference [Электронный ресурс].— URL: https://docs.microsoft.com/ru-ru/sql/dmx/data-mining-extensions-dmx-reference?view=sql-server-ver15 (Дата обращения 16.05.2020). Загл. с экр. Яз. рус.

9 *Фуше, Г.* Pro SQL Server 2008 Analysis Services / Г.Фуше— Apress, 2009.— С.480

10 Создание прогнозов [Электронный ресурс].— URL: https://docs.microsoft.com/ru-ru/sql/tutorials/creating-predictions-basic-data-mining-tutorial?view=sql-server-2014 (Дата обращения 16.05.2020). Загл. с экр. Яз. рус.