Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и
информационных технологий

Разработка информационной технологии Business Intelligence средствами RapidMiner

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студента 4 курса 421 группы
направления 09.03.01 «Информатика и вычислительная техника»
факультета компьютерных наук и информационных технологий
Ивлиева Алексея Сергеевича

Научный руководитель
к. ф.-м.н., д. экон.н., профессор _____ Л.В. Кальянов
дата, подпись

Заведующий кафедрой
к. ф.-м.н., доцент _____ Л.Б. Тяпаев
дата, подпись

Саратов 2018 год

ВВЕДЕНИЕ

Деятельность современного предприятия связана с анализом большого объёма данных, как внутренней, так и внешней информации. Реализовать анализ такой информации возможно только средствами современных информационных технологий. Оной из таких технологий, является Bussines Intelligence.

Таким образом персонал не может обойтись без специализированных программных средств, способствующих анализу и расчёту данных в предоставленных отчётах предприятия.

Целью дипломной работы является реализация информационной технологии Business Intelligence средствами Rapid Miner для предприятия, являющимся продуктовым супермаркетом. На основе данных о продажах предприятия и критериях оценки риска реализаций, провести анализ данных, построить графики на основании получившихся данных, провести поиск скрытых закономерностей на основе метода линейной регрессии, оптимизировать данные предприятия следующим образом, чтобы группы наименований не находились в критических зонах, минимизировать расходы на нереализованную продукцию.

В главе 1 «Теоретические основы Business Intelligence (BI)» содержится общая информация о BI и произведено сравнение с аналитикой и бизнес – аналитикой. Подробно рассмотрена архитектура построения технологии BI, приведены примеры её реализации в различных сферах. Описаны методы и технологии, используемые в BI, такие как : нейронные сети, дерево принятия решений, правила индукции, технология OLAP – кубов, технология ETL. Так же в данной главе рассмотрено специальное программное обеспечение Rapid Miner, предназначенное для анализа данных и обоснование того, почему был сделан выбор на данное программное обеспечение.

В главе 2 « Разработка процесса минимизации убытков предприятия, на основе анализа данных, средствами Rapid Miner» описан процесс реализации технологии BI для предприятия, являющимся продуктовым магазином, в

программном обеспечении Rapid Miner. В Rapid Miner построены графики, для поиска скрытых закономерностей, с использованием метода линейной регрессии. Проведена настройка работы алгоритма BI в автоматизированном режиме, с целью организовать итерационную работу бухгалтера в процессе формирования отчётности. На основе построенного алгоритма, проведён анализ данных – отчёта предприятия, для подтверждения полезности и истинности технологии BI, с целью минимизации расходов на нереализованную продукцию.

В настоящее время существует большое количество предприятий, для которых необходима реализация технологии BI, в связи с необученностью персонала проводить анализ в данной сфере или невозможностью обработать информацию вручную, из – за того, что данные для анализа имеют слишком большой объем информации.

Основное содержание работы

В главе 1 **« Теоретические основы Business Intelligence (BI)»** инструменты для анализа данных, построения отчетов и запросов могут помочь пользователям преодолеть большое количество данных для того, чтобы синтезировать из них значимую информацию, сегодня эти инструменты в совокупности попадают в категорию, называемую Business Intelligence. [1]

Принято три типа аналитики, которые в совокупности составляют весь BI, это:

1) Описательный метод.
2) Предиктивный метод.
3) Предписывающий метод.

Цели и методологии, используемые для каждого из трех типов аналитики, различаются. Именно эти различия отличают аналитику от BI.

Аналитика сосредоточена на создании полезной информации из источников данных, BI использует аналитику для создания улучшенной измеримой деловой активности. [8]

Аналитика может содержать любой из трех методов, когда BI включает в себя все три одновременно, для создания новых, уникальных методов принятия решений. [12]

При сортировке данных в качестве первого шага Business Intelligence , используется четыре установленных классификации измерений, такие как :

1) Категориальные Данные
2) Порядковые Данные
3) Данные Интервалов
4) Данные по Коэффициенту

Три основных компонента: описательный, прогнозирующий и предписывающий, в процессе BI, может помочь фирме найти возможности в данных, которые прогнозируют будущие возможности, и помощь в выборе курса действий, чтобы максимизировать стоимость и производительность. [2]

В главе 1.2 « **Архитектура системы Business Intelligence»** основным назначением BI-систем является обеспечение возможности анализа больших объемов информации для решения бизнес – задач. [13]

В системе BI, должны выполняться следующие условия :

1) Эффективное получение данных.
2) Обработка данных.
3) Предоставление данных конечным пользователям.

Процесс выполнения работы Business Intelligence следует по строгому алгоритму, в котором:

1) Источник данных, в последующем используемый для анализа.

   Может содержать следующие типы данных:

   a) CSV таблицы.
   b) Базы данных.
   c) PDF файлы, с дополнительной проверкой.
   d) Exel таблицы.
2) Процесс Business Intelligence.
3) Набор данных.
4) Моделирование.

   Включает в себя такие типы, как :

   a) Схемы.
   b) Кубы.
   c) Метаданные.
5) Сервер Business Intelligence.
6) В итоге отчёт, состоящий из различных графиков и текстовых документов. [3]

Успешность архитектуры проверяется на правиле: при штатной работе системы, пользователи получают именно ту информацию, которая им нужна и именно тогда, когда она нужна. [15]

При построении архитектуры Business Intelligence, используются два разных подхода :

1) Системы с интегрированными детальными данными – процесс создания более длительный, но система будет стабильна.

2) Создание специализированных витрин данных без интеграции всех данных в детальном слое – более быстрое решение, которое не будет так эффективно и устойчиво к изменениям в источниках. [9]

В главе 1.3 **«Технологии Business Intelligence (BI)»** в используемые средства BI для решения поставленной проблемы, встроены такие инструменты, как: OLAP, нейронные сети, деревья принятия решений, правила индукции, логистическая регрессия, дискриминантный анализ, средства моделирования, визуализация, ETL. [4]

Нейронные сети используются в задачах классификации (где выход является категориальной переменной) или для регрессий (где выходная переменная непрерывна). [10]

Дерево принятия решений — средство поддержки принятия решений, использующееся в статистике и анализе данных для прогнозных моделей. Структура дерева представляет собой «листья» и «ветки». [14]

Правила индукции – метод получения набора правил для классификации случаев.

Логистическая используется для прогнозирования двоичных переменных со значениями, такими как yes, no или 0, 1 и иногда переменными класса. [5]

OLAP — набор технологий для обработки информации, включающих динамическое построение отчётов, анализ данных, прогнозирование ключевых показателей бизнеса. [11]

ETL — один из основных процессов в управлении хранилищами данных, который включает в себя: извлечение данных из внешних источников, их трансформация и очистка. [6]

В главе 1.4 **«Обоснование выбора Rapid Miner»**, Rapid Miner определена, как мощная и многопользовательская платформа, она служит для создания, передачи и обслуживания наукоемких данных.

Используются методы текстового майнинга, веб – майнинга, автоматической тональности анализа интернет – форумов, также доступны анализ и прогнозирование временных рядов. [7]

Rapid Miner позволяет использовать сложные визуализации, такие как 3-D графики, матрицы рассеяния и самоорганизующиеся карты.

Платформа Rapid Miner в настоящее время имеет как платную лицензию, так и бесплатную с ограничениями, в сравнении с конкурентами где имеется только платная версия.

Плюсы Rapid Miner в сравнении с конкурентами :

1. Хороший GUI**.** Каждый функциональный блок собран в кубик. Например, в SPSS Modeler 50 узлов, в RapidMiner 250 в базовой загрузке.

2. Хорошие инструменты подготовки данных. К примеру SPSS возможностей гораздо меньше.

3. Расширяемость. Используется язык программирования R.

4. Присутствует совместимость с Hadoop

5. Архитектурно данные снаружи.

7. Сервер умеет сразу строить минимальные отчёты.

Если сравнивать Rapid Miner с другими программами, то у Rapid Miner гораздо шире функциональные возможности по обработке, больше узлов.

В главе 2 **«Разработка процесса минимизации убытков предприятия, на основе анализа данных, средствами RapidMiner»**, файл с данным для обработки находится в формате .CSV.

Учитываются такие пункты как: Наименование товара, коэффициент постоянной нужды в товаре, коэффициент стоимости данного продукта, коэффициент заинтересованности в покупке, коэффициент уровня рекламы бренда, коэффициент крайнего срока годности до момента его реализации, коэффициент дальности расположения продукта от кассового аппарата. Все пункты, кроме раздела наименования продукта, являются целым числом. Каждый атрибут, помимо наименования товара, содержит показатель от нуля до десяти.

В данной работе реализована технология ETL, где данные проходят следующий алгоритм:

1) Преобразование структуры.
2) Агрегирование.
3) Перевод значений.
4) Создание новых данных.
5) Очистка данных.
6) Загрузка данных на сервер и последующая их обработка.

Производится обязательная настройка данных в Rapid Miner перед их обработкой. Настройка кодировки windows – 1251 и метод разделения столбцов пользовательский.

Для первой строки установлено значение Name, так как они будут являться заголовками атрибутов. Первый раздел устанавливаем в значение polynominal, ячейки в значение numeric.

Использованы такие операторы как Remove Duplicates и Multiply, предназначенные в последующей обработки данных удалять дубликаты и возможность создавать несколько процессов с одним входным файлом.

В главе 2.2 **«Первичный анализ данных средствами Rapid Miner»** на данном этапе построенного алгоритма BI, использованы такие операторы, как :

1) Read CSV
2) Multiply
3) Remove Duplicates
4) Generate attributes
5) Select attributes
6) Sort
7) Generate Log
8) Filter Examples

После первичного прохождения алгоритма, составили диаграмму соотношения коэффициента срока годности к коэффициенту стоимости

продукта и два графика гибкости всех и одной группы наименований, встроенными средствами Rapid Miner «Plot View» на основе получившихся данных, для поиска скрытых закономерностей.

В главе 2.3 **«Вторичный анализ данных, поиск общего значения»** на данном этапе построенного алгоритма BI, дополнительно использованы такие операторы, как :

1) Generate Empty Attributes
2) Generate Empty Attributes (1)
3) Generate Empty Attributes (2)
4) Map
5) Multiply (2)

Для вычисления общего коэффициента, создана функция подсчёта, описанная в операторе Generate Attributes (2) в разделе Function expressions.

Оператором Select Attributes произвели поиск атрибутов и их вывод с условием игнорирования остальных атрибутов.

Оператором Sort реализовали сортировку в данных с условием increasing по атрибуту общего коэффициента.

Вычисления общего коэффициента производится по формуле:

$$\frac{Кб+Кс+Кз+Кр+Кср+Крт}{6}$$

После прохождения алгоритма в результате получили вывод групп наименований, находящимися в критических зонах реализаций, то есть имеющие коэффициент ниже 5.0, составили данное заключение на графике нерентабельных наименований, с целью упрощенного восприятия данного результата для пользователя.

В главе 2.4 **«Оптимизация продаж»** с целью минимизировать расходы на нереализованную продукцию добавили атрибут Коэффициент расположения на витрине.

В данной работе магазин имеет четыре витрины, нумерация происходит с нижней витрины – 1, до верхней – 4.

Так как витрины с наименованием два и три находятся в большем приоритете, коэффициент витрины под номером три будет варьироваться в значение от восьми до десяти, коэффициент витрины под номером два будет находиться в пределе от пяти до семи. Витрины под наименованием один и четыре будут иметь значения ниже пяти.

На конечном этапе построенного алгоритма BI, дополнительно использованы такие операторы, как :

1) Guess Types
2) Generate Attributes (3)

Произвели перезапись данных через оператор Generate Empty Attributes для автоматического добавления к существующим данным раздела коэффициента расположения на полке. Зарезервировали ячейки Integer.

Через оператор Map написали автоматическое распределение каждого наименования в своё ключевое место с условиями True, False.

Провели новую обработку данных с перезаписанной формулой вычисления, описанной в операторе Generate Attributes (3) :

$$\frac{Кб+Кс+Кз+Кр+Кср+Крт+Крп}{7}$$

Оператором Guess Types изменили формат атрибута с Nominal на Integer для более точного отображения результата. Так же возможной ошибки одновременной обработки данных в форматах Numeric и Nominal.

После прохождения алгоритма, в результате получили обновленные текстовые данные, и составленную диаграмму анализа вероятности продаж.

В результате с использованием технологии BI, наименования, находившиеся в группе с наименьшим показателем, из этой группы показатели стали более благоприятные для их реализации. Некоторые наименования, такие как Молоко вид 3, Кофе вид 4 и вовсе вышли из зоны риска реализации, благодаря наилучшему расположению.

При этом все остальные наименования в наименьшей группе стали иметь общий коэффициент не менее четырех от пяти нормализированных.

Благодаря верному анализу и распределению наименований, расход данной организации сократиться на 20 – 30 процентов, в сравнении с организацией, для которой не применена данная технология.

ЗАКЛЮЧЕНИЕ

В ходе дипломной работы построен алгоритм, реализующий технологию Bussines Intelligence в программном обеспечении RapidMiner.

Решены задачи:

1) Поиск скрытых закономерностей методом линейной регрессии и на их основе выполнены дальнейшие условия.
2) Расчёт общего коэффициента каждого наименования.
3) Организация наиболее продуктивных показателей предприятия, методом прохождения алгоритма.
4) Построение графиков, в целях обнаружения критических групп наименований.
5) Стимуляция продаж наименований, имеющих низкий показатель привлекательности, методом верного распределения коэффициента расположения товара на витрине.
6) Изменение и перезапись данных от неактуальной информации, содержащейся в входном файле.

Список использованных источников

1 Larissa T. Moss, Shaku Atre, Business Intelligence Roadmap. Addison Wesley. 2003. С. 230 – 270.

2 Ranjit Bose, Vijayan Sugumaran. Application of intelligent agent technology for Managerial Data Analysis and Mining. University of New Mexico. 1999. С. 77 – 90.

3 Mahesh Raisinghani, Business Intelligence in the Digital Economy: Opportunities, Limitations, and Risks. Idea Group Publishing. 2004. С. 45 – 91.

4 Zbigniew Michalewicz, Martin Schmidt, Matthew Michalewicz, Constantin Chiriac, Adaptive Business Intelligence. Springer. 1998. С. 132 – 178.

5 Carlos Soares, Yonghong Peng, Jun Meng, Takashi Washio, Zhi-Hua Zhou, Applications of Data Mining in E-Business and Finance. IOS Press. 2008. С. 30 – 54.

6 А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод, Методы и модели анализа данных: OLAP и Data Mining. 2006. Издательство bhv. С. 88 – 112.

7 Davide Moraschi, Business Intelligence with MicroStrategy Cookbook. Packt Publishing Enterprise. 2013. С. 215 – 281.

8 Stephane Tuffery, Data Mining and Statistics for Decision Making. Wiley publication. 2011. С. 504 – 573.

9 Michael J.A. Berry Gordon S. Linoff, Data Mining Techniques For Marketing, Sales, and Customer Relationship Management, Second edition. Wiley publishing, Inc. 2004. С. 122 – 147.

10 Antti Syväjärvi, Jari Stenvall, Data Mining in Public and Private Sectors: Organizational and Government Applications. Published in the United States of America by Information Science Reference (an imprint of IGI Global). 2010. С. 212 – 237.

11 Ken Withee, Microsoft® Business Intelligence FOR DUMmIES. Wiley publishing, Inc. 2010. С. 387 – 451.

12 Robert Stackowiak, Joseph Rayman, Rick Greenwald, Oracle® Data Warehousing and Business Intelligence Solutions. Wiley publishing, Inc. 2007. 215 – 274.

13 С.Д. Ильенкова, Л.М. Гохберг, В.И. Кузнецов, С.Ю. Ягудин, Инновационный менеджмент. Московский международный институт эконометрики, информатики, финансов и права. 2003. С. 4 – 60.

14 Ю.Ф. Тельнов, Интеллектуальные информационные системы (учебное пособие). Московский международный институт эконометрики, информатики, финансов и права. 2003. С. 4 – 110.

15 Алёхина Г.В., Информационные технологии в экономике и управлении. Московский международный институт эконометрики, информатики, финансов и права. 2003. С. 40 – 65.