На правах рукописи

Михалькевич Илья Сергеевич

ИНСТРУМЕНТЫ И МЕТОДЫ АНАЛИЗА СЛАБОСТРУКТУРИРОВАННЫХ ДАННЫХ В ОПТИМИЗАЦИИ МАРКЕТИНГОВЫХ КОММУНИКАЦИЙ

08.00.13 – Математические и инструментальные методы экономики

Автореферат
диссертации на соискание ученой степени
кандидата экономических наук

Москва – 2018

Работа выполнена на кафедре «Прикладная информатика» ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации»

Научный руководитель:	доктор экономических наук, доцент **Лукьянов Павел Борисович**
Официальные оппоненты:	**Уринцов Аркадий Ильич,** доктор экономических наук, профессор, ФГОБУ ВО «Российский экономический университет имени Г.В. Плеханова», заведующий кафедрой управления информационными системами и программирования
	Корсун Олег Николаевич, доктор технических наук, профессор, государственный научно-исследовательский институт авиационных систем, начальник лаборатории
Ведущая организация:	**Федеральное государственное автономное образовательное учреждение высшего образования «Волгоградский государственный университет»**

Защита состоится 4 апреля 2018 года в 10-00 часов на заседании диссертационного совета Д 505.001.03 на базе ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации» по адресу: Ленинградский проспект, д. 55, ауд. 213, Москва, ГСП-3, 125993.

С диссертацией можно ознакомиться в диссертационном зале Библиотечно-информационного комплекса ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации» по адресу: Ленинградский проспект, д. 49, комн. 203, Москва, ГСП-3, 125993 и на официальном сайте ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации»: http://www.fa.ru.

Автореферат разослан 29 января 2018 г.

Ученый секретарь
диссертационного совета Д 505.001.03,
доктор экономических наук, профессор Герасимова Елена Борисовна

I ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Одним из ключевых факторов роста любой компании является повышение эффективности управления внутренними ресурсами. Одним из таких ресурсов является клиентская база компании, потенциал которой в значительной степени реализуется за счёт маркетинговых коммуникаций. Высокие стандарты качества, предъявляемые к клиентской базе, а также знания, полученные в результате её глубокого анализа, являются необходимыми условиями эффективного управления маркетинговыми коммуникациями.

В условиях наличия больших массивов разнородных данных у организации их анализ значительно затрудняется. Существующие механизмы интеграции систем не позволяют полностью учесть специфику бизнес-процессов организации: это чревато потерей важной информации, необходимой для принятия решений. Кроме этого, в силу особенностей процесса накопления данных назревает потребность в использовании специфических средств, методов анализа и моделирования.

Анализ опыта российских и зарубежных компаний показывает, что применение маркетинговых коммуникаций в бизнесе позволяет значительно повысить лояльность существующих клиентов, привлечь большое количество новых клиентов и, таким образом, увеличить продажи. Использование такого дорогостоящего инструмента маркетинговых коммуникаций, как телефонный звонок, требует оптимизации использования этого канала коммуникации.

Лизинг автомобилей является динамично развивающейся отраслью на российском рынке финансовых услуг, однако, применение маркетинговых коммуникаций в лизинге имеет весьма ограниченный характер. Всё больше компаний, как российских, так и зарубежных понимают важность комплексного подхода к повышению качества клиентской базы и повышению эффективности маркетинговых коммуникаций. Поэтому разработка соответствующих инструментов и методов является актуальным направлением научного исследования.

Степень разработанности темы исследования. С целью совершенствования алгоритмов сопоставления информации о клиентах в работе изучены методы дедупликации и извлечения текста на основе нечёткого поиска, так как именно в них затрагиваются наиболее сложные проблемы работы с данными в значительной степени

субъективного характера. Фундаментальные исследования в данном научном направлении освещены в работах Р. Хемминга, Ф. Дамерау и В.И. Левенштейна. В этих работах предлагается ввести метрику для оценки расстояния между строковыми последовательностями. Дальнейшие работы, расширяющие область применения предложенных метрик, были посвящены, в основном, так называемым оффлайн алгоритмам нечёткого поиска, таких как нечёткий поиск с индексацией и алгоритм расширения выборки. В работах Л.М. Бойцова вводится понятие сигнатуры, отражающей наличие тех или иных символов алфавита в строке, затем предлагается алгоритм индексирования хеш-таблиц, построенных на сигнатуре. Работы Е. Укконена посвящены нечёткому поиску методом N-грамм. Достоинством этого подхода является простота реализации и хорошая производительность алгоритмов, недостатком – то, что очень близкие друг к другу строковые последовательности могут оказаться незамеченными. Помимо перечисленных методов существует множество интересных, но малоэффективных вариантов, таких как, фонетические алгоритмы, один из которых впервые предложен Р. Расселом, или адаптация префиксных деревьев к задачам нечёткого поиска Т.Г. Меррета.

Из представленных метрических алгоритмов нечёткого поиска наиболее адекватные результаты показывают алгоритмы на основе метрики Дамерау-Левенштейна. Однако, понятие метрики Дамерау-Левенштейна можно расширить с целью реализации тонкой настройки нечёткого поиска, а также поиска последовательностей заданного типа.

Оценку качества дедупликации данных часто проводят либо с точки зрения скорости работы алгоритмов дедупликации, либо путём непосредственного применения метрик строкового расстояния. Однако, остаётся нереализованной потребность в оценке качества дедупликации независимо от выбранных метрик и скорости работы алгоритмов.

Основы теории оптимизации поиска и сортировки данных разработаны в фундаментальных работах Д. Кнута, Э. Дейкстры и развиты в исследованиях Н. Вирта, Г. Гарсией-Молины, Дж. Ульмана, Дж. Видом. Вместе с тем почти не представлено работ по оптимизации поиска по составному ключу.

Другим инструментом, позволяющим повысить эффективность маркетинговых коммуникаций, является анализ данных, машинное обучение и математическое

моделирование отклика на маркетинговые коммуникации. В области машинного обучения разработано огромное количество различных методик, таких как методы кластеризации (графовые: связных компонент, кратчайшего незамкнутого пути, ForEl; статистические: EM-алгоритм, k-средних; иерархические), решающие деревья, регрессия, нейронные сети. Большинство методов рассмотрены в работах К.В. Воронцова, описаны их достоинства и недостатки. Различные методы визуализации многомерных данных изучены в работах К. Пирсона, К. Карунена, М. Лоэва, Т. Кохонена, А. Зиновьева и других учёных. Тем не менее, вопросы отображения многомерных данных, имеющих признаки с высоким коэффициентом эксцесса требуют отдельного рассмотрения.

Популярным методом в анализе маркетинговых коммуникаций является RFM-анализ (recency – давность, frequency – частота, monetary – затраченные деньги), теоретические основы которого развиты в работах П. Фадера, Б. Харди и К. Ли. Эти зарубежные исследования описывают наиболее общие характеристики в поведении клиентов. Вместе с тем существуют перспективы уточнения набора значимых признаков с точки зрения оптимизации маркетинговых коммуникаций в условиях современного российского бизнеса.

Существующие подходы к моделированию конверсии маркетинговых коммуникаций часто не отражают специфики прикладных областей, где необходимо применять перечисленные методы. Это, в свою очередь, требует развития математических моделей отклика на маркетинговые коммуникации, учитывающих особенности клиентской базы лизинговой компании, а также методов нечёткого поиска информации в слабоструктурированных или неструктурированных текстовых данных.

Объект исследования. Объектом исследования является компания, специализирующаяся на лизинге автомобилей для юридических лиц.

Предмет исследования. Предметом исследования является моделирование отклика на маркетинговые кампании, построение портрета клиента и его оценка, а также методы обработки данных, включая их очистку и стандартизацию.

Область исследования. Тема работы соответствует направлениям исследований, описанным в разделах 1.4. «Разработка и исследование моделей и математических методов анализа микроэкономических процессов и систем: отраслей народного хозяйства, фирм и предприятий, домашних хозяйств, рынков, механизмов

формирования спроса и потребления, способов количественной оценки предпринимательских рисков и обоснования инвестиционных решений»; 2.6. «Развитие теоретических основ методологии и инструментария проектирования, разработки и сопровождения информационных систем субъектов экономической деятельности: методы формализованного представления предметной области, программные средства, базы данных, корпоративные хранилища данных, базы знаний, коммуникационные технологии» Паспорта научной специальности" 08.00.13 – «Математические и инструментальные методы экономики» (экономические науки).

Целью исследования является разработка инструментов и методов анализа слабоструктурированных данных клиентской базы лизинговой компании и построение модели отклика клиентов на маркетинговые коммуникации.

Для достижения цели исследования были сформулированы и решены следующие **задачи**:

1. Применение системного подхода к таргетированию маркетинговых коммуникаций посредством обогащения данных о клиентах за счёт внутренних ресурсов клиентской базы.
2. Прогнозирование отклика клиентов на маркетинговые коммуникации, позволяющее значительно повысить лояльность клиентов и увеличить продажи.
3. Адаптация метода нечёткого поиска для обнаружения данных заданного типа в строковой последовательности.
4. Реализация алгоритма дедупликации данных на основе методов нечёткого сравнения символьных строк.
5. Разработка метода предварительного анализа признаков для включения в регрессионную модель отклика на маркетинговые коммуникации.
6. Выявление значимых признаков клиентов на основе анализа их влияния на результат маркетинговых коммуникаций с помощью разработанного метода.
7. Выявление значимых признаков клиентов с использованием нелинейных методов отображения данных с последующей кластеризацией.
8. Разработка алгоритма нормирования признаков для включения в регрессионную модель с учётом их распределения в выборке.
9. Построение регрессионной модели отклика на маркетинговые коммуникации с использованием полученных признаков.

10. Реализация полученных методов и моделей в программном комплексе и внедрение в производственный процесс.

Методология и методы исследования. Основу работы составили теоретические и методические разработки по анализу данных в сфере маркетинга, математической статистики, эконометрического моделирования, теории информации и компьютерной лингвистики, теории баз данных, проектирования информационных систем.

В ходе исследования применялись методы системного анализа, экспертных оценок, статистического и сравнительного анализа, а также табличные и графические приёмы визуализации данных.

Информационной базой исследования послужили данные российских и зарубежных аналитических агентств, информация из открытых источников о внедрении систем управления клиентской базой и оптимизации маркетинговых коммуникаций, научные труды российских и зарубежных исследователей, данные российской компании «Европлан», специализирующейся на лизинге автомобилей и спецтехники.

Научная новизна исследования заключается в модификации и развитии методов работы с клиентской базой компании, построении новой модели отклика клиентов на маркетинговые коммуникации. Основные результаты, содержащие элементы научной новизны:

1. Предложено и обосновано использование преобразования метрики Дамерау-Левенштейна для поиска дублирующихся текстовых записей и вычленения текстовых данных заданного типа.

2. Предложены методы оценки качества дедупликации и извлечения данных.

3. Предложено и обосновано использование модификации составного ключа с пустыми значениями в базе данных для возможности индексированного поиска новых объектов.

4. Предложен и обоснован метод нормализации многомерных признаков для корректировки отображения многомерных данных на нелинейное многообразие, вложенное в пространство большей размерности.

5. Выявлены новые значимые факторы, позволяющие уточнить оценку вероятности отклика клиента на маркетинговые коммуникации.

6. Построена регрессионная модель для оценки вероятности отклика клиента на маркетинговые коммуникации с использованием выявленных факторов.

Положения, выносимые на защиту:

1. Предложено и обосновано преобразование метрики Дамерау-Левенштейна, что расширяет возможности данной метрики при поиске дублирующихся записей и вычленении текстовых данных заданного типа (С. 59-62).

2. Разработаны методы оценки качества дедупликации и извлечения данных о клиентах на основе дополнительной информации, имеющейся в базе данных, а также полученной в результате эксперимента (С. 63-65).

3. Предложена модификация составного ключа с пустыми значениями в базе данных и обосновано её использование при индексированном поиске новых объектов (С. 69-72).

4. Предложен метод сегментации клиентской базы с помощью отображения многомерных данных на нелинейное многообразие, вложенное в пространство большей размерности, с дальнейшей кластеризацией этих данных (С. 83-85).

5. На основе анализа результатов взаимодействия компании с клиентами выявлены новые значимые характеристики клиентов, позволяющие уточнить оценку вероятности их отклика на маркетинговые коммуникации (С. 74-76, 78-79, 82, 86).

6. Построена регрессионная модель оценки вероятности отклика клиента на маркетинговые коммуникации, учитывающая влияние выявленных значимых характеристик клиента (С. 87-96).

7. Разработан программный комплекс автоматизации исходящих маркетинговых коммуникаций, позволивший существенно увеличить доход лизинговой компании (С. 97-122).

Теоретическая значимость представленных в работе результатов состоит в развитии методов оптимизации маркетинговых коммуникаций, математических моделей отклика клиентов на маркетинговые коммуникации и инструментов обработки клиентской базы. Основные положения и выводы диссертации дополняют существующие методы анализа слабоструктурированных данных и прогнозирования поведения клиентов. Материалы и обобщения, полученные в диссертации, могут служить теоретической основой для дальнейшего развития областей исследования, связанных с управлением клиентской базой компании и оптимизацией маркетинговых коммуникаций.

Практическая значимость исследования. Разработанные в диссертации методы и модели ориентированы на широкое использование в крупных и средних организациях, имеющих значительную клиентскую базу. Применение разработанных методов и моделей обеспечивает экономический эффект, выражающийся в виде увеличения продаж, повышения лояльности клиентов и наращивания клиентского портфеля.

Самостоятельную практическую значимость имеют:

1. Повышение эффективности маркетинговых коммуникаций за счёт увеличения конверсии при сохранении среднего чека, а, следовательно, увеличение прибыли и окупаемости инвестиций в маркетинговые коммуникации.

2. Предотвращение оттока клиентов в результате соблюдения контактной политики.

3. Получение представления об основных факторах (времени, прошедшего с последней покупки, количестве и стоимости покупок, источнике обращения клиента, и др.) и характере их влияния на конверсию в маркетинговых коммуникациях лизинговой компании.

4. Внедрение программного комплекса, позволившего:

- повысить скорость формирования списка клиентов для проведения маркетинговых коммуникаций за счёт очистки и структуризации данных;
- устранить неопределённость, обеспечить полноту, точность и согласованность данных в управленческой отчётности благодаря преобразованию исходных данных адаптированными методами.

5. Сокращение трудовых затрат на создание управленческой отчётности о результатах маркетинговых коммуникаций.

Для достижения полученных практических результатов были применены:

- Адаптированный метод нечёткого поиска для обнаружения данных заданного типа в строковой последовательности.
- Алгоритм нормирования признаков для включения в регрессионную модель с учётом их распределения в выборке.
- Метод оценки качества дедупликации и извлечения данных.
- Регрессионная модель для оценки вероятности отклика клиента на маркетинговые коммуникации, учитывающая влияние выявленных значимых характеристик клиента.

Результаты исследования нашли практическое применение в маркетинговой деятельности лизинговой компании ПАО «Европлан» и используются в учебном процессе ФГОБУ ВО «Финансовый Университет при Правительстве Российской Федерации» в преподавании дисциплины «Технологии интеллектуального анализа данных».

Степень достоверности, апробация и внедрение результатов исследования. Достоверность полученных результатов была подтверждена большим фактическим материалом, результаты исследования согласуются с фундаментальными положениями экономической теории. Методика проведения расчётов соответствует критериям, предъявляемым к научному подходу, и позволяет получить объективные результаты. Разработка программных средств велась с использованием современных платформ и языков программирования.

Результаты исследования обсуждались и получили положительные отзывы на межвузовских и международных научно-практических конференциях: на межвузовском круглом столе «Молодые учёные о проблемах отечественной науки» (Москва, Финансовый университет, 21 апреля 2014 г.), на научной конференции «Научные достижения молодых исследователей» (Москва, Финансовый университет, 29 марта 2014 г.), на IV международном конкурсе научных работ аспирантов и студентов (Москва, Финансовый университет, 28 апреля 2015 г.), на московской научно-практической конференции «Студенческая наука» (Москва, Финансовый университет, 30 ноября 2015 г.), на V международном конкурсе научных работ аспирантов и студентов (Москва, Финансовый университет, 4 апреля 2016 г.), на международной научно-практической конференции «Актуальные проблемы развития современной науки и образования» (Москва, Научное издательство «Ар-Консалт», 30 апреля 2016 г.).

Материалы диссертации используются в практической деятельности Управления исследований ПАО «Европлан». По материалам исследования внедрен программно-аппаратный комплекс, предназначенный для решения задач маркетинговых коммуникаций, в том числе: объединения данных о клиенте компании из внутренних систем и внешних источников в единую сущность, централизованного хранения данных об истории взаимодействия с клиентом, внедрения математических моделей поведения клиента для повышения эффективности коммуникаций, формирования аналитической

отчетности. Выводы и основные положения диссертации дают эффект в виде увеличения конверсии маркетинговых коммуникаций с 6,6 до 12,3% и, таким образом, получения дополнительной прибыли в размере 9 000 тыс. руб. ежегодно.

Материалы диссертации используются в учебном процессе кафедрой «Прикладная информатика» ФГОБУ ВО «Финансовый Университет при Правительстве Российской Федерации» в преподавании учебной дисциплины «Технологии интеллектуального анализа данных».

Результаты внедрения подтверждены соответствующими документами.

Публикации по теме диссертации. По теме диссертации опубликовано 6 работ общим объемом 3,39 п.л. (весь объем авторский), в том числе 4 работы авторским объемом 2,5 п.л. опубликованы в рецензируемых научных изданиях, определенных ВАК при Минобрнауки России.

Структура диссертации определена целью, задачами и логикой исследования и состоит из введения, пяти глав, заключения, списка литературы из 161 источника и 1 приложения. Работа изложена на 141 странице и содержит 52 рисунка, 26 таблиц, 80 формул.

II ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

1. Проведен анализ исследований, посвященных поиску способов привлечения новых клиентов и удержания имеющихся клиентов с помощью маркетинговых коммуникаций

По данным исследований ведущих компаний в отрасли, развитие каналов маркетинговых коммуникаций позволяет значительно увеличить продажи, а внедрение систем управления клиентской базой позволяет не только повысить лояльность клиентов, но и, в конечном счете, обеспечивает значительный рост прибыли от прямых продаж.

Цель маркетинговых коммуникаций – привлечь наибольшее количество новых клиентов и удержать максимальное количество существующих клиентов. Особенностью задачи является то, что она носит массовый характер. Маркетинговые коммуникации нацелены на широкую аудиторию и не включают в себя уникальные предложения для ключевых клиентов.

Целесообразность проведения и модернизации маркетинговых коммуникаций иллюстрирует статистика различных аналитических агентств в Таблица 1.

Таблица 1 – Исследования маркетинговых коммуникаций

Экономические показатели
- американские компании в результате низкого качества маркетинговых коммуникаций теряют 41 млн. долл. ежегодно; - доходы американских компаний от email-маркетинга достигли 156 млн. долл. в 2015 году; - расходы компаний на email-маркетинг растут на 10 процентов ежегодно; - средняя окупаемость инвестиций в email-маркетинг составляет 44,25 доллара на каждый потраченный доллар; - 56 процентов опрошенных бизнесменов планируют увеличить расходы на email-маркетинг.
Исследования лояльности
- 69 процентов клиентов, оставшихся довольными взаимодействием с компанией, рекомендуют её своим друзьям и партнёрам по бизнесу; - 58 процентов клиентов, оставшихся недовольными взаимодействием с компанией, больше никогда не воспользуются услугами этой компании; - в среднем лояльный клиент тратит в 10 раз больше стоимости первоначальной покупки.
Исследования каналов коммуникаций
- 80 процентов получателей email-рассылок открывают письма, пришедшие от компаний; - 44 процентов получателей email-рассылок за год делают хотя бы одну покупку в следствие получения рекламного сообщения; - клиенты, подписанные на email-рассылки, совокупно тратят больше на 83 процента, размеры их заказов больше на 44 процента, а частота покупок выше на 28 процентов; - персонализированные рассылки увеличивают частоту кликов на 14 процентов и конверсию на 10 процентов; - 70 процентов получателей купонов и скидок по почте используют их в течение следующей недели; - 27 процентов потребителей часто упоминают о том, что их любимым компаниям следует больше инвестировать в коммуникации по электронной почте; - 39 процентов маркетологов не имеют никакой стратегии для мобильной электронной почты; - до сих пор 68 процентов коммуникаций обслуживаются посредством телефонных звонков; - отсутствие обратной связи в коммуникациях через социальные сети вызывает ускорение оттока клиентов до 15 процентов.

Источник: составлено автором.

Показано, что объём продаж и лояльность клиентов в значительной степени определяет качество маркетинговых коммуникаций. Описанные закономерности делают очевидной целесообразность внедрения систем управления клиентской базой,

оптимизации маркетинговых коммуникаций и инвестиций в прогнозирование поведения клиентов. По прогнозу Transparency Market Research рынок предикативной аналитики вырастет до 6,5 миллиардов долларов к 2019 г.

Системы управления клиентской базой и прогностические модели маркетинговых коммуникаций всё чаще внедряются в средних и больших российских организациях. В 2014 году компания Форексис передала в промышленную эксплуатацию банку Траст систему, позволяющую прогнозировать поведение клиентов и проводить оптимизацию целевого маркетинга. Компания Human Labs реализовала многочисленные проекты по очистке и стандартизации данных у таких заказчиков, как Сбербанк, ВТБ24, Ситибанк, Ренессанс, Почта России, ИКЕА и др.

Проблемы управления клиентской базой и оптимизации маркетинговых коммуникаций также стали предметом изучения у ряда российских исследователей (Андреева А.В., Агаджанов А.А., Траньков Д.О., Афонина К.В., Волкова Н.В., Бургат В.В., Свиридова О.И.).

2. Предложено преобразование метрики Дамерау-Левенштейна для поиска дублирующихся записей и вычленения текстовых данных заданного типа

В диссертации рассмотрены и систематизированы исследования российских и зарубежных авторов (В.И. Левенштейна, Л.М. Бойцова, Дёмина И.С., Р. Хемминга, Ф. Дамерау, Е. Укконена, Т.Г. Меррета и др.) по вопросам управления качеством данных, раскрывающие проблемы нечёткого поиска по тексту и дедупликации данных. Также кратко проанализированы такие методы нечёткого сравнения строк, как двоичные сигнатуры и метод N-грамм. Показаны ограничения существующих методов при распознавании текстовых данных заданного типа.

Для реализации нечёткого сравнения текстовых данных определяется метрика, которая отражает степень сходства этих данных. Наиболее подходящей метрикой для текстовых строк, введённых вручную является метрика Левенштейна, и, её усовершенствованный вариант – метрика Дамерау-Левенштейна, отражающая расстояние редактирования или замены символа в тексте. Алгоритм, измеряющий расстояния между строками m и n имеет временную сложность O(m*n) и требует O(m*n) памяти.

Метрика Дамерау-Левенштейна – расстояние между двумя конечными строковыми последовательностями, соответствующее минимальному количеству

операций вставки, удаления, замены и транспозиции соседних символов, необходимому для преобразования одной строковой последовательности в другую.

Предлагается адаптация метрики Дамерау-Левенштейна для поиска текста заданного типа. Например, имея запись «Поляков Василий Викторович, тел.: +7(985)345-55-32, родился 26.08.1974, Советская, д.5» необходимо распознать телефон клиента. Разработан алгоритм, который с наибольшей (субъективной) точностью позволяет распознавать символьные подпоследовательности и относить их к тому или иному классу. Для этого предложено использовать метрику Дамерау-Левенштейна с использованием матрицы стоимостей замены символов, представленной в таблице 2.

Таблица 2 – Матрица стоимостей замены символов. \d – любая цифра, \w – любая буква

	\d	\w	-	(	)	.	+
\d	0	2	1	1	1	1	1
\w	2	0	1	1	1	1	1
-	2	1	0	1	1	2	1
(	1	1	1	0	1	2	1
)	1	1	1	1	0	2	1
.	1	1	2	2	2	0	2
+	1	1	1	1	1	2	0

Источник: составлено автором.

После чего необходимо составить по меньшей мере одной эталонной последовательности в качестве представителя каждого класса.

Представители класса телефонов: [+7(985)-665-43-55], [8(985)6654355], [9856654355]. Представители класса дат: [2015-01-07], [9.01.1993].

Матрица стоимостей замены символов позволяет лучше отделить последовательности, для которых характерны какие-то определённые символы, от всех остальных. Аналогичным образом можно ввести веса удаления и вставки малозначимых, или, наоборот, очень важных символов.

С введением матрицы стоимостей сложность алгоритма несколько увеличится за счёт поиска значений в матрице (1):

$$O'(m * n * \log k) \quad (1)$$

где m и n – размеры строковых последовательностей, k – размер матрицы стоимостей, $2\log_2 k$ – максимальное количество операций для поиска стоимости замены пары символов k1, k2.

Применение метрики потребует выбрать порог отсечения, для которого представители классов будут считаться несопоставимыми.

Далее показано, что применение такого подхода позволяет улучшить нечёткий поиск относительно введённых автором критериев качества дедупликации и извлечения данных.

3. Предложены методы оценки качества дедупликации и извлечения данных о клиентах

В виду отсутствия достаточной и репрезентативной информации о том, какие записи на самом деле являются дублирующимися, оценка качества дедупликации на основе исходных данных возможна только при участии человека. Для получения объективной оценки вводится коэффициент качества дедупликации, который отражает отношение вероятностей нахождения общих контактных данных клиентов из одной группы и их нахождения у клиентов из разных групп по формулам (2), (3), (4), (5):

$$DQ = \frac{\sum_{i=1}^{n} P_{same}(x_i)}{1 + \sum_{i=1}^{n} P_{diff}(x_i) * n} \quad (2)$$

$$P_{same}(x) = \frac{\sum_{j=1}^{l} sign(x = x_j)}{l - 1}, x \in G_k, x_j \in G_k \backslash x \quad (3)$$

$$P_{diff}(x) = \frac{\sum_{j=1}^{n-l} sign(x = x_j)}{n - l}, x \in G_k, x_j \notin G_k \quad (4)$$

$$sign(x = x_j) = \begin{cases} 1, x = x_j \\ 0, x \neq x_j \end{cases} \quad (5)$$

где n – число объектов, l – число объектов в группе G_k, k – количество групп, $P_{same}(x)$ – вероятность наличия общих контактных данных у объекта x с объектами группы G_k, $P_{diff}(x)$ – вероятность наличия общих контактных данных у объекта x с объектами других групп.

Введён показатель качества извлечения данных, заданный отношением (6) – это отношение количества данных, извлечённых верно, к количеству данных, извлечённых неверно:

$$EQ = \frac{TP}{1 + FP} \quad (6)$$

Показатель качества дедупликации позволяет дать независимую оценку дедупликации записей о клиентах при наличии дополнительной информации, например, контактных данных. Определить правильность извлечения контактных данных можно с помощью эксперимента, например, используя эти данные в маркетинговых коммуникациях.

Статистика извлечения и проверки телефонных номеров отражена в таблице 3.

Таблица 3 – Статистика извлечения и проверки телефонных номеров

Метод	**Номеров извлечено**	**Номеров прозвонено выборочно**	**Актуальных номеров в выборке**	**Оценка количества актуальных номеров**
Метрика Дамерау-Левенштейна с матрицей замены символов	1 169 211	10000	2 424	283 445
Метрика Дамерау-Левенштейна	1 344 885	10000	2 000	268 977
Метрика Левенштейна	1 630 072	10000	1 525	248 655
Регулярные выражения	1 253 670	10000	2 100	263 271

Источник: составлено автором.

В таблице 4 приведены показатели качества дедупликации и извлечения данных с применением различных методов.

Таблица 4 – Показатели качества дедупликации и извлечения данных

Метод	**DQ**	**EQ**
Метрика Дамерау-Левенштейна с матрицей замены символов	6.11	0.32
Метрика Дамерау-Левенштейна	6.04	0.25
Метрика Левенштейна	5.98	0.18
Регулярные выражения	4.66	0.21

Источник: составлено автором.

4. Предложена модификация составного ключа с пустыми значениями в базе данных для индексированного поиска новых объектов

Поиск сущностей в базе данных можно ускорить за счёт применения индексов. Время, требуемое для соединения таблиц без индексированного поиска, вычисляется по формуле (7):

$$T = t * \prod N_i \quad (7)$$

где t – время, требуемое на выбор одной записи, N_i – количество записей в таблице i.

В то время, как соединение этих же таблиц с применением индексированного поиска потребует значительно меньшего количества времени, вычисляемого по формуле (8):

$$T = t * \prod \prod \log_2 N_i \tag{8}$$

Если в двух таблицах A, B имеются записи об одних и тех же сущностях, а их идентификаторами служат A_id, B_id соответственно, то связи между этими сущностями можно представить, как в таблице 5:

Таблица 5 – Пример таблицы связки сущностей

A_id	B_id
1	1761
2	NULL
3	3487
4	4217
NULL	5544
NULL	6111
5	7444

Источник: составлено автором.

Замена NULL в целочисленном поле A_id разложением на MAX(A_id)+1 и 2^31-1 и B_id на MAX(B_id)+1 и 2^31-1 соответственно позволит быстрее найти как старую, так и новую запись, примеры которых записаны в таблицах 6 и 7.

Таблица 6 – Пример найденной старой записи

A_id	B_id
NULL	6111

Источник: составлено автором.

Таблица 7 – Пример найденной новой записи

A_id	B_id
6	6111

Источник: составлено автором.

Таким образом, в рамках решения задачи об индексации строк с частичной потерей данных был предложен алгоритм, позволяющий ускорить поиск таких строк. Предложенный алгоритм можно применять при составлении справочников, для анализа данных, для измерения качественных характеристик клиентской базы и прочих задач.

5. Предложена нормализация многомерных признаков клиентов для корректировки их отображения на нелинейное многообразие, вложенное в пространство большей размерности

В диссертации приведены результаты кластеризации клиентской базы лизинговой компании с применением карт Кохонена и упругих карт. В данном разделе предложена нормализация признаков клиентов для их корректного отображения на этих картах. Нормализация позволяет дать больший контраст между значениями признака там, где сосредоточена большая часть данных. В работе даётся формальное определение такой нормализации и приводится в формуле (9):

$$X_i \to X_i' | x_{ij}' = x_{F(x_{ij})}, j \in n \tag{9}$$

где n – объём выборки, $x_{i1}, \dots, x_{in}$ – множество значений признака X_i, $x_{i1}', \dots, x_{in}'$ – множество значений признака X_i', $F(x_{ij})$ – эмпирическая функция распределения X_i задаётся формулой (10):

$$F(x_{ij}) = \overline{P(x < x_{ij})} = P(x < x_{ij}) + \frac{P(x = x_{ij})}{2} \tag{10}$$

$x_{F(x_{ij})}$ – квантиль уровня $F(x_{ij})$ для стандартного нормального распределения задаётся формулой (11):

$$F(x) = N = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \tag{11}$$

Достоинством такого подхода является простота реализации, недостатком – необходимость хранения таблицы преобразования возможных значений по интервалам.

Выходом из этой ситуации является подбор известной функции распределения для заданного признака по принципу максимального правдоподобия. В работе вводится минимизируемый функционал, заданный отношением (12):

$$F(x_{ij}) - F_k(x_{ij}) \to min \tag{12}$$

Для описания распределения признаков «выручка» и «количество автомобилей в парке», подходит семейство гамма-распределений, описываемое выражением (13):

$$\Gamma(k,\theta) = \frac{\gamma(x,\frac{x}{\theta})}{\Gamma(k)\theta^k} \quad (13)$$

Требуется минимизировать функционал, заданный отношением (14)

$$F(x_{ij}) - \frac{\gamma(x_{ij},\frac{x_{ij}}{\theta})}{\Gamma(k)\theta^k} \to min \quad (14)$$

поиском параметров θ и k, например, при помощи метода обобщённого приведённого градиента, реализованного в Microsoft Office Excel, или с помощью других численных методов.

Достоинством подхода с выбором известного распределения является наличие отображения в виде формулы, недостатком – возможные отклонения значений теоретической функции распределения от значений эмпирической функции распределения.

6. Выявлены новые значимые характеристики клиентов, позволяющие уточнить оценку вероятности их отклика на маркетинговые коммуникации

В диссертации рассмотрены работы, посвящённые RFM-анализу (Ван Де Поль, Артур Хьюз, Дэвид Шепард, Гарри Рассел, Джон Миглауч и др.) и методам расчёта жизненной ценности клиента (Брюс Харди, Вильям Кан, Натаниэль Лин и др.). RFM-анализ опирается на наиболее общие экономические и психологические закономерности, отражающие поведение клиента, а именно: давность покупки (recency), частота покупок (frequency) и количество затраченных денег (monetary). В контексте решаемой в исследовании задачи приведённые постулаты были уточнены, а также были выявлены другие важные закономерности.

Конверсия маркетинговых коммуникаций по классическому определению – отношение числа успешных попыток коммуникаций к общему числу коммуникаций.

Для анализа независимого влияния характеристик клиентов на конверсию вводится коэффициент характерности признака, который численно выражает типичность признака для целевого клиента по сравнению со всей группой. При этом используется формула расчёта конверсии (15):

$$Conv_i = \frac{TS_i}{TS_i + TF_i} \quad (15)$$

где $Conv_i$ – конверсия в группе с i-м признаком, TS_i – количество удачных попыток коммуникаций по группе, TF_i – количество неудачных попыток коммуникации по группе.

Тогда коэффициент характерности признака рассчитывается по формуле (16):

$$C_i = \frac{S_i - P_i}{P_i} = \frac{\frac{p_i}{TS} - \frac{V_i}{V}}{\frac{V_i}{V}} = \frac{V \times p_i}{V_i \times TS} - 1 \quad (16)$$

где C_i – коэффициент характерности признака i у целевой группы, отражающий ценность обладания информацией о факте наличия признака i, P_i – доля клиентов с признаком i, V_i – число клиентов с признаком i, V – общее число клиентов в группе, S_i – доля клиентов с признаком i в целевой группе, p_i – число клиентов в целевой группе с признаком i

По итогам анализа результатов маркетинговой кампании был получен портрет идеального лизингополучателя, который описывается утверждением «клиент с выручкой от 8,5 до 34 млн, пришедший по рекомендации дилера, который часто делает покупки и недавно проявлял интерес к очередному предложению». Такой портрет может оказать помощь в стратегическом планировании и полезен при создании новых маркетинговых кампаний.

Применение карты Кохонена дало похожий результат. Следующие признаки оказались наиболее информативными: «Самостоятельно проявлял интерес к покупке» (выражено в группе с высокой конверсией), «Скорость покупок» (в группе с низкой конверсией выражается низкими значениями), «Проявлял ли интерес к различным транспортным средствам» (в группе с низкой конверсией выражено отрицательно), «Обратился через дилера» (в группе с высокой конверсией выражено отрицательно), «Упрощённая система налогообложения» (явно указывает на группу с низкой конверсией), «Прошло времени с последней покупки» (в группах с высокой конверсией высокие значения не представлены).

Применение упругих карт позволяет оценить количество кластеров и найти наиболее близкий к интуитивному вариант разбиения (если алгоритм кластеризации сходится к нескольким вариантам). На базе упругих карт была выполнена кластеризация

методом k-means с 5 кластерами и составлена статистика кластеров. Результаты кластеризации представлены в таблице 8.

Наиболее высокая конверсия (8,26%) была представлена в кластере №1, содержащем 16,33% испытаний, с характерными признаками: «Скорее интересовался разными ТС», «Недавно что-либо покупал», «Самостоятельно интересовался покупкой».

Таблица 8 – Статистика кластеров после применения упругой карты

Cluster #	Avg Y, %	Count, %	DiffType	FromLastBuy	SelfProject	RTaxType	FromDealer	Tspeed
1	8,26	16,33	0,97	-0,83	0,92	-0,33	-0,28	-0,18
2	5,52	15,39	-1,03	-0,7	0,48	-0,11	-0,55	-0,48
3	0,89	39,31	-0,45	0,6	-0,58	0,24	0,27	-0,67
4	5,54	17,35	0,61	-0,08	0,06	-0,09	0,89	1,29
5	5,64	11,62	0,59	0,17	-0,05	-0,07	-1,12	1,22

Источник: составлено автором.

Наиболее низкая конверсия (0,89%) представлена в кластере №3, содержащем 39,31% испытаний, с характерными признаками: «Скорее не интересуется разными ТС», «В последнее время ничего не покупал», «Никогда самостоятельно не интересовался покупкой», «Редко совершает покупки».

Применение упругих карт позволило целенаправленно улучшить кластеризацию, чего не удалось достичь с применением карт Кохонена: значительно лучше были классифицированы клиенты с наиболее низкой конверсией, относительно хорошо были классифицированы высококонверсионные клиенты. Наглядные результаты были представлены на графиках плотности распределения данных, распределения кластеров и значимых признаков. Эта информация может быть использована для построения регрессионных моделей и планирования маркетинговых коммуникаций.

7. Построена регрессионная модель оценки вероятности отклика клиента на маркетинговые коммуникации, учитывающая влияние выявленных значимых характеристик клиента

Другой ключевой задачей в повышении эффективности маркетинговых коммуникаций является прогнозирование отклика клиентов. Для этого была построена логистическая регрессионная модель отклика клиентов на маркетинговые коммуникации с учётом выявленных признаков. Логистическая модель хороша тем, что задаёт непрерывную функцию вероятности по вектору признаков.

По критерию Фишера самыми значимыми оказались признаки «Дней с последней покупки» (Z-value = 19,02), «Количество покупок» (Z-value = 8,39), «Клиент интересовался разными транспортными средствами» (Z-value = 4.16).

Для оценки классификационной мощности модели была построена ROC-кривая. При наихудшем сценарии (случайная классификация) значение показателя AUC равно 0.5. Значение AUC, равное 0,87 можно считать признаком хорошей классификации. Например, при пороге отсечения, дающем 15% вероятности отклика клиентов на маркетинговые коммуникации, в сегменте, отсечённом моделью, вероятность отклика не превышает 1.5%.

С учётом специфики экономической задачи признак «Количество покупок» оказался несколько более значимым, чем «Количество затраченных денег», что показано в таблице 9.

Таблица 9 – Z-статистика для признаков «Количество покупок» и «Количество затраченных денег» в моделях

predictor	z-value
TransactionCount	8.386
TransactedPrice	5.470

Источник: составлено автором.

Тем не менее, эти признаки имеют очень высокую корреляцию: 0,807 в выборке и 0,889 – в генеральной совокупности. Поэтому, в дальнейшем был использован только один из этих признаков.

8. Применение полученных результатов в автоматизации маркетинговой деятельности лизинговой компании

С целью внедрения полученных результатов в маркетинговую деятельность лизинговой компании разработан программный комплекс, в котором реализованы следующие функции:

- очистка и стандартизация данных о клиентах;
- преобразование и нормализация слабоструктурированных данных;
- дедупликация сущностей – записей о клиентах (юридических лицах);
- формирование выборки клиентов для осуществления маркетинговых коммуникаций на основе построенной регрессионной модели.

Условно программный комплекс можно разделить на следующие модули:

- Модуль нечёткого сравнения идентификационных данных

– Модуль сбора, очистки и стандартизации данных

– Модуль преобразования характеристических данных

– Модуль объединения данных

– Модуль принятия решения о загрузке контрагента для коммуникации в рамках маркетинговой компании (МК)

– Модуль загрузки контрагентов для коммуникации в рамках МК

На выходе образуется база данных объединённых дублей контрагентов, информация о которых непротиворечива и обогащена из внешних источников. Эта база данных готова для формирования управленческой отчётности и автоматизации процесса управления маркетинговыми коммуникациями.

Произведено тестирование маркетинговых коммуникаций с использованием внедрённого комплекса и без него:

- основной результат – повышение конверсии маркетинговых коммуникаций с 6,6 до 12,3% (в 1,86 раза);
- значительно повышена скорость загрузки контрагентов: с 10-40 минут до 15-45 секунд (на 1 МК);
- повышена лояльность клиентов за счёт более полного соблюдения контактной политики;
- предотвращён рост оттока клиентов, пришедших в ранние периоды (доля клиентов, пришедших 5 лет назад, и сделавших покупку в целевой период осталась на уровне 12%), а по новым клиентам зафиксировано увеличение доли клиентов, совершивших повторную покупку (доли клиентов, пришедших 3 и 1,5 года назад, и сделавших покупку в целевой период возросли с 15 до 16% и с 22 до 24% соответственно);
- создана платформа для формирования оперативной отчётности и мониторинга;
- расчётное увеличение прибыли компании за счёт внедрения комплекса:

$$МЧИ * 12 * СДИ * (1 - 1 / 1{,}86) = 17\,566\,489 * 12 * 9{,}2\% * 0{,}4624 = 8\,967\,510 \text{ руб.}$$

где МЧИ – среднемесячные чистые инвестиции; СДИ – средний доход на инвестицию;

- затраты на внедрение комплекса: 4 000 000 руб;
- временные затраты: 1 год (поэтапное внедрение. Февраль 2014 – Февраль 2015).

III ЗАКЛЮЧЕНИЕ

В ходе проведённого исследования были получены следующие результаты:

1. Предложено и обосновано использование преобразования метрики Дамерау-Левенштейна для поиска дублирующихся текстовых записей и вычленения текстовых данных заданного типа.
2. Предложены и протестированы методы оценки качества дедупликации и извлечения данных.
3. Предложена и использована модификация составного ключа с пустыми значениями в базе данных для возможности индексированного поиска новых объектов.
4. Предложен метод сегментации клиентской базы с помощью отображения многомерных данных на нелинейное многообразие, вложенное в пространство большей размерности, и дальнейшей кластеризации.
5. Выявлены новые значимые факторы, позволяющие уточнить оценку вероятности отклика клиента на маркетинговые коммуникации.
6. Построена регрессионная модель для оценки вероятности отклика клиента на маркетинговые коммуникации с использованием выявленных факторов.
7. Разработан программный комплекс, работающий на регулярной основе, предназначенный для сбора разнородных данных из разных источников, дедупликации записей о клиентах, преобразования и сведения данных в единую структуру.
8. Автоматизирован процесс маркетинговых коммуникаций, с имплементацией регрессионной модели (и алгоритма сегментации).
9. Использование усовершенствованных методов позволило оптимизировать загрузку ресурсов маркетинговых коммуникаций. В результате было достигнуто значительное увеличение конверсии (с 6,6 до 12,3%) маркетинговых коммуникаций при сохранении среднего чека на прежнем уровне. Таким образом, ежегодная прибыль компании от исходящих маркетинговых коммуникаций была увеличена в 1,86 раза.
10. Устранено негативное влияние коммуникаций на лояльность клиентов, количество жалоб на повторные звонки и рассылки сведено к нулю благодаря соблюдению контактной политики. Предотвращён рост оттока клиентов.

IV СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в рецензируемых научных журналах и изданиях, определенных ВАК при Минобрнауки России:

1. Михалькевич, И.С. Преобразование метрики Дамерау-Левенштейна для обработки данных, используемых в маркетинговых коммуникациях / И.С Михалькевич // Микроэкономика. – 2015. – №5. – С. 101-106. (0,62 п.л.).

2. Михалькевич, И.С. Моделирование конверсии маркетинговых коммуникаций в лизинговой компании / И.С. Михалькевич // Инновации и инвестиции. – 2016. – №2. – С. 165-169. (0,6 п.л.).

3. Михалькевич, И.С. Анализ маркетинговых коммуникаций с помощью нелинейных методов отображения данных / И.С. Михалькевич // Фундаментальные исследования. – 2016. – № 4-1. – С. 201-207. (0,68 п.л.).

4. Михалькевич, И.С. Методы повышения эффективности обработки клиентской базы данных / И.С. Михалькевич // Инновации и инвестиции. – 2016. – №5. – С. 104-108. (0,6 п.л.).

Статьи в других научных изданиях:

5. Михалькевич, И.С. Повышение достоверности слабоформализованных данных / И.С. Михалькевич // Научные записки молодых исследователей. – 2014. – №2. – С. 17-21. (0,64 п.л.).

6. Михалькевич, И.С. Управление качеством корпоративных данных, предназначенных для автоматической обработки [Электронный ресурс] / И.С. Михалькевич // Электронный научный журнал. – 2016. – №4. – С. 564-567. (0,25 п.л.). Режим доступа: http://co2b.ru/docs/enj.2016.04.pdf (дата обращения: 30.10.2017).

Подписано в печать 10.01.2018
Объем 1,5 усл.п.л.
Тираж 120 экз. Заказ № 206
Отпечатано в типографии «Реглет»
г. Москва, пр-т Мира, д.38
+7(495)979-98-99, www.reglet.ru