Федеральное государственное образовательное бюджетное учреждение высшего образования
«Финансовый университет при Правительстве Российской Федерации»

На правах рукописи

Михалькевич Илья Сергеевич

ИНСТРУМЕНТЫ И МЕТОДЫ АНАЛИЗА СЛАБОСТРУКТУРИРОВАННЫХ ДАННЫХ В ОПТИМИЗАЦИИ МАРКЕТИНГОВЫХ КОММУНИКАЦИЙ

Специальность 08.00.13 – Математические и инструментальные методы экономики

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата экономических наук

Научный руководитель:
доктор экономических наук, доцент
Лукьянов Павел Борисович

Москва – 2016

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

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 [138]. Высокие стандарты качества, предъявляемые к клиентской базе, а также знания, полученные в результате её глубокого анализа, являются необходимыми условиями эффективного управления маркетинговыми коммуникациями.

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 [129, 75, 130, 121, 70]. 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 [50, 107]. Поэтому разработка соответствующих инструментов и методов является актуальным направлением научного исследования.

Степень разработанности темы исследования. Одним из инструментов повышения эффективности маркетинговых коммуникаций является поддержание качества клиентской базы на высоком уровне. В

клиентской базе высокого качества информация о клиентах обладает следующими свойствами:

- полнота: информация присутствует в системе в необходимом для обеспечения бизнес-процессов количестве;
- непротиворечивость: в системе отсутствуют взаимоисключающие факты;
- структурированность: в системе однозначно определён способ получения информации.

Для обеспечения соответствия информации данному критерию высокого качества применяются комплексные подходы к её преобразованию, в состав которых могут входить следующие процедуры:

1. Очистка (приведение данных к заданному формату). Это могут быть простые процедуры замены значений в соответствии с заданными правилами. Для обработки текстовых значений наиболее популярным методом являются регулярные выражения.

2. Стандартизация (приведение данных в соответствие набору допустимых значений). Данная процедура слабо поддаётся автоматизации, особенно это касается приведения справочников из разных систем к единому стандарту.

3. Нормализация (организация данных в БД посредством создания таблиц и отношений, обеспечивая устранение избыточности и несогласованных зависимостей). Здесь, в частности, применяются методы обработки слабоструктурированных данных. Для обработки слабоструктурированных данных известных форматов используются синтаксические парсеры (от англ. parse – «разбор», «структурный анализ»). В случае несоответствия данных формату или отсутствия форматирования применяются регулярные выражения и методы нечёткого поиска по тексту.

4. Дедупликация (объединение дублирующихся записей на основе формальных критериев дупликации, а также объединение связанных атрибутов и сущностей в соответствии с установленными правилами). Здесь

также могут быть использованы регулярные выражения и методы нечёткого сравнения символьных строк, основанные на описанных ниже метриках.

Важным является рассмотрение исследований, развивающих методы дедупликации записей на основе нечёткого поиска, а также методы нечёткого поиска по тексту записей заданного типа, так как именно в них изучаются наиболее сложные проблемы работы с данными, которые носят в большой степени субъективный характер.

Фундаментальные исследования в данном научном направлении освещены в работах Р. Хемминга, Ф. Дамерау и В. Левенштейна [60, 78, 101]. 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. Аналогично автор предлагает алгоритм индексирования хеш-таблиц, построенных на сигнатуре [11].

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. Кроме того, понятие метрики Дамерау-Левенштейна

можно расширить и открыть новые возможности применяемых на её основе алгоритмов.

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 [67, 73, 77, 80, 95, 102, 105, 106, 108, 124, 126, 131, 133, 136, 148, 151, 158]. Большинство методов рассмотрены в работах К.В. Воронцова, описаны их достоинства и недостатки [16].

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 [86, 87]. Эти зарубежные исследования описывают наиболее общие характеристики в поведении клиентов.

Существующие современные подходы часто не отражают специфики прикладных областей, где необходимо применять перечисленные методы, что в свою очередь требует их развития для решения практических задач.

В проведённом исследовании предлагается развитие математических моделей отклика на маркетинговые коммуникации лизинговой компании, а также методов нечёткого поиска информации в слабоструктурированных или неструктурированных текстовых данных на базе метрики Дамерау-Левенштейна.

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, описанных в пунктах 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).

Цель проводимого исследования – повысить эффективность исходящих маркетинговых коммуникаций лизинговой компании посредством разработки инструментов и методов обработки и анализа слабоструктурированных данных.

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

1. Разработан системный подход к таргетированию маркетинговых коммуникаций в условиях ограниченных финансовых ресурсов.
2. Разработан подход к обогащению данных о клиентах за счёт внутренних ресурсов клиентской базы, позволяющий значительно повысить лояльность.
3. Произведена адаптация метода нечёткого поиска для обнаружения данных заданного типа в строковой последовательности.

4. Реализован алгоритм дедупликации данных на основе методов нечёткого сравнения символьных строк.

5. Разработан метод предварительного анализа признаков для включения в регрессионную модель отклика на маркетинговые коммуникации.

6. Выполнен анализ влияния признаков на результат маркетинговых коммуникаций с помощью разработанного метода. Выявлены значимые по введённому коэффициенту характерности признаки.

7. Выполнен анализ влияния признаков на результат маркетинговых коммуникаций с помощью нелинейных методов отображения данных. Проведено сравнение результатов с результатами, полученными в п. 6.

8. Разработан алгоритм нормирования признаков для включения в регрессионную модель с учётом их распределения в выборке.

9. Построена регрессионная модель отклика на маркетинговые коммуникации. Выявлены значимые признаки по критерию Фишера. Проведено сравнение результатов с результатами, полученными в пп. 6 и 7, а также с классической теорией.

10. Разработан и внедрен в эксплуатацию программный комплекс, автоматизирующий следующие процессы:

- 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.

Методология и методы исследования. Основу работы составили теоретические и методические разработки учёных и специалистов 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 подхода, статистического и сравнительного анализа, экспертных оценок, а также табличные и графические приёмы визуализации данных.

Информационная база исследования. Информационной базой исследования послужили данные российской компании, специализирующейся на лизинге автомобилей и спецтехники, данные российских и зарубежных аналитических агентств, данные открытых источников о внедрении систем управления клиентской базой и оптимизации маркетинговых коммуникаций, научные труды российских и зарубежных исследователей, акты федерального законодательства Российской Федерации.

Научная новизна диссертационного исследования заключается в следующем:

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. Для нечёткого сопоставления дублирующихся записей о клиентах и вычленения их контактных данных из текстовых строк автором введена матрица стоимостей замены символов, позволяющая тонко настроить нечёткий поиск и лучше различать в тексте данные разных типов с помощью метрики Дамерау-Левенштейна (с. 59-62).

2. Автором предложены методы оценки качества дедупликации и извлечения данных о клиентах на основе анализа косвенных признаков, а также на основе выборочной проверки (с. 63-65).

3. Автором предложен и обоснован алгоритм модификации составного ключа для ускорения поиска в базе данных записей о клиентах из разных систем (с. 69-72).

4. Автором предложено и обосновано использование нормализации многомерных признаков, позволяющей сделать их отображение на двумерное нелинейное многообразие более информативным и упростить процесс кластеризации клиентской базы (с. 83-85).

5. Автором проанализированы факторы, влияющие на вероятность отклика клиентов в маркетинговых коммуникациях. Классическая модель, основанная на RFM-анализе, была дополнена применительно к решаемой экономической задаче (с. 74-76, 78-79, 82, 86).

6. Автором построена логистическая регрессионная модель отклика клиентов на маркетинговые коммуникации, в которую были включены изученные факторы. Это позволило усилить классифицирующую способность модели (с. 87-96).

7. Построенная модель, разработанные алгоритмы дедупликации и извлечения данных были внедрены при разработке программного комплекса автоматизации маркетинговых коммуникаций лизинговой компании, что в

совокупности позволило значительно увеличить доход компании от исходящих маркетинговых коммуникаций (с. 97-122).

Теоретическая значимость представленных в работе результатов состоит в разработке подходов к оптимизации маркетинговых коммуникаций, развитии математических моделей отклика на маркетинговые коммуникации и инструментов обработки клиентской базы.

Практическая значимость работы состоит в

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

2. Предотвращении оттока клиентов в результате соблюдения контактной политики.

3. Получении 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. Внедрении 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. Сокращении трудовых затрат на создание управленческой отчётности о результатах маркетинговых коммуникаций.

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 VII Международном студенческом конгрессе (Москва, Финансовый университет, 29 марта 2014 г.), на межвузовском круглом столе «Молодые учёные о проблемах отечественной науки» (Москва, Финансовый университет, 21 апреля 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 кафедрой «Прикладная информатика» Финансового университета в преподавании учебной дисциплины «Технологии интеллектуального анализа данных».

Результаты внедрения подтверждены соответствующими документами.

Публикации. 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.

ГЛАВА 1
ИССЛЕДОВАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ. ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ ОБРАБОТКИ И АНАЛИЗА ДАННЫХ

1.1 Маркетинговые коммуникации. Обзор предметной области

В силу развития корпоративных телекоммуникационных технологий, появления новых каналов коммуникаций, совершенствования маркетинговых инструментов, накопления большого количества данных о клиентах, маркетинговые коммуникации представляют собой самостоятельный интерес для исследования.

Маркетинговые коммуникации – в широком смысле, сообщения, с помощью которых осуществляется взаимодействие с рынком [70].

Предметом диссертационного исследования являются исходящие маркетинговые коммуникации, для которых дадим несколько более узкое определение:

Исходящие маркетинговые коммуникации – коммуникации с клиентом, осуществляемые по инициативе компании с задействованием таких маркетинговых инструментов, как интернет, e-mail, sms, прямые продажи (исходящие звонки).

Для осуществления исходящих маркетинговых коммуникаций необходимо развивать соответствующие каналы продаж. Для осуществления возможности проведения и контроля централизованных исходящих маркетинговых коммуникаций необходимо полностью контролировать входящие и исходящие потоки информации в каналах продаж, а, следовательно, содержать эти каналы на собственные средства компании. Содержание таких дорогостоящих каналов коммуникаций, как колл-центр, доступно не всем лизинговым компаниям, однако, за счёт эффекта от

масштаба, наличия большой клиентской базы и эффективного осуществления маркетинговых коммуникаций содержание таких каналов может быть выгодным.

По данным исследований [129, 75, 130, 121, 70] 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 счёте, обеспечивает значительный рост прибыли от прямых продаж.

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

Вышеописанные тенденции 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 млрд. долл. к 2019 г. [135].

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 [50]. Внедрение системы позволило:

- повысить процент отклика на предложения;
- повысить доходность маркетинговых кампаний;
- увеличить лояльность клиентов и стоимость клиентской базы;
- повысить эффективность каналов коммуникаций.

Проблемы управления клиентской базой и оптимизации маркетинговых коммуникаций также стали предметом изучения у ряда российских исследователей [1, 6, 7, 12, 15, 43, 48].

Таблица 1 – Исследования маркетинговых коммуникаций

Экономические показатели	
Ежегодные потери американских компаний в результате низкого качества маркетинговых коммуникаций, млн. долл.	41
Доходы американских компаний от email-маркетинга в 2015 году, млн. долл.	156
Ежегодный рост расходов компаний на email-маркетинг	0,1
Средняя окупаемость инвестиций в email-маркетинг	44,25
Доля опрошенных бизнесменов, планирующих увеличить расходы на email-маркетинг	0,56
Исследования лояльности	
Доля клиентов, рекомендующих её своим друзьям и партнёрам по бизнесу, среди оставшихся довольными взаимодействием с компанией	0,69
Доля, клиентов которые больше никогда не воспользуются услугами компании, среди оставшихся недовольными	0,58
Превышение трат среднестатистического лояльного клиента над стоимостью его первоначальной покупки	10
Исследования каналов коммуникаций	
Доля получателей email-рассылок открывающих письма от компаний	0,8
Доля получателей email-рассылок делающих хотя бы одну покупку за год в следствие получения рекламного сообщения	0,44
Увеличение трат клиентов, подписавшихся на email-рассылки	0,83
Увеличение размеров заказов клиентов, подписавшихся на email-рассылки	0,44
Увеличение частоты покупок клиентов, подписавшихся на email-рассылки	0,28
Повышение частоты кликов в персонализированных рассылках	0,14
Повышение конверсии от персонализированных рассылок	0,1
Доля получателей купонов и скидок использующих их в течение следующей недели	0,7
Доля потребителей, часто упоминающих о том, что их любимым компаниям следует больше инвестировать в коммуникации по электронной почте	0,27
Доля маркетологов, не имеющих никакой стратегии для мобильной электронной почты	0,39
Доля коммуникаций, обсуживающихся посредством телефонных звонков	0,68
Ускорение оттока клиентов из-за отсутствия обратной связи в коммуникациях через социальные сети	0,15

Источник: составлено автором на основе данных [75, 70, 76, 122, 81, 83, 150, 90, 85, 118, 159, 155, 96, 81].

Компания Human Labs реализовала многочисленные проекты по очистке и стандартизации данных у таких заказчиков, как Сбербанк, ВТБ24, Ситибанк, Ренессанс, Почта России, ИКЕА и др. [107]. Это позволило им:

- ускорить подготовку и проведение маркетинговых коммуникаций;
- адекватно оценить стоимость клиентской базы;
- сохранить лояльность своих клиентов и предотвратить отток;
- ускорить процессы интеграции внутренних систем.

Маркетинг лизинговых услуг предназначен для удовлетворения потребностей бизнеса в части финансирования основных средств. Маркетинг лизинговых услуг ставит следующие цели:

- достижение максимального уровня потребления услуги лизинга;
- достижение максимальной удовлетворённости потребителя-лизингополучателя;
- максимальное расширение возможностей лизингополучателя для ведения бизнеса;
- максимально широкое представление выбора программ финансирования.

Завьялов П.С. в своей работе подразделяет маркетинг лизинговых услуг на 4 основных компонента [23, С. 441]:

- услуга лизинга, включающая в себя программы лизинга, товарную политику, ассортимент продуктов;
- цена – денежная сумма, взимаемая лизингодателем за конкретный товар или услугу;
- сбыт – доведение товара или услуги до потребителя. Каналами продаж могут выступать различные автодилеры, поставщики спецтехники и пр.;
- комплекс маркетинговых коммуникаций, реализуемый четырьмя средствами воздействия: личной продажей, рекламой, стимулированием сбыта, связями с общественностью.

Каждое из перечисленных воздействий представляет собой набор маркетинговых инструментов. Воздействие на предпочтения целевой аудитории является основной задачей маркетинговых коммуникаций [114].

Для продвижения своих услуг лизинговые компании широко используют различные маркетинговые инструменты:

- звонки потенциальным («холодные» звонки) и текущим клиентам по имеющимся базам данных;
- звонок-рассылка-звонок. Выполняется: звонок клиенту, email-рассылка с коммерческим предложением, повторный звонок с консультацией;
- рассылки на электронную почту рекламного характера;
- разработка и поддержка сайта в сети Интернет.

Поведение потенциального лизингополучателя, связанное с принятием решения о покупке, подвержено влиянию следующих факторов [43]:

- время, требуемое на прохождение полного цикла бизнес-процессов, начиная от озвучивания клиентом потребности, подачи документов, заканчивая передачу имущества в лизинг;
- уровень сервиса, оказываемый клиенту на протяжении всего цикла бизнес-процессов;
- цена сопутствующих необходимых услуг, таких, как страхование;
- наличие дополнительных услуг.

Эти факторы необходимо учитывать, выстраивая политику маркетинговых коммуникаций.

В работах зарубежных исследователей в качестве инструмента комплекса маркетинга лизинговых услуг анализируется целевой маркетинг [109, 127], который, в отличие от массового, предполагает разработку лизинговых программ под каждый из сегментов рынка. Предельным случаем целевого маркетинга является прямой маркетинг. В концепции прямого маркетинга коммуникации с клиентом осуществляются индивидуально [28]. Его инструментами могут быть:

- телефон;
- электронная почта;
- классическая почта;
- Интернет;
- личная встреча.

Преимущества прямого маркетинга для лизингополучателя:

- максимальное соответствие потребностям;
- экономия времени;
- удобство.

В качестве иной формы маркетинговых коммуникаций выделяют интерактивный маркетинг, осуществляемый посредством взаимодействия через сайт в сети Интернет [50]. Данный вид коммуникаций позволяет дистанционно оставить заявку, задать вопрос, сделать оперативный отчёт, получить доступ к большому объёму информационных ресурсов.

Интерактивные маркетинговые коммуникации позволяют лизингодателю:

- повысить лояльность клиентов;
- сократить расходы на обслуживание входящего потока обращений.

1.2 Моделирование в маркетинге. Математический аппарат

Имея большую базу клиентов, делать маркетинговые объявления для всех сразу дорого и не эффективно. Необходимо найти таких клиентов, которые с наибольшей вероятностью откликнутся на маркетинговое предложение, иначе говоря, *целевую аудиторию*. Такая процедура называется – *таргетированный маркетинг*. Основными методами таргетированного маркетинга являются *сегментирование* и *скоринг*.

В зарубежных исследованиях приводятся наиболее ранние определения понятия «сегментирование рынка» и «скоринг» [125, 144].

Сегментирование рынка – представляет собой процесс выявления групп потребителей с различными характеристиками спроса для диверсификации маркетинговых усилий и тонкой настройки продукта в соответствии с требованиями рынка.

Скоринг – определение вероятности положительного результата взаимодействия с клиентом на основе скоринговой модели.

К инструментам сегментирования относятся деревья решений, карты Кохонена, нейронные сети, кластерный анализ. К инструментам скоринга относятся логистическая регрессия и ROC-анализ.

1.2.1 Деревья решений

Деревья решений удобно использовать при описании портрета покупающего клиента. Этот инструмент предиктивной аналитики достаточно хорошо изучен и применяется в различных прикладных областях [66, 67, 79, 92, 103, 107, 111, 126].

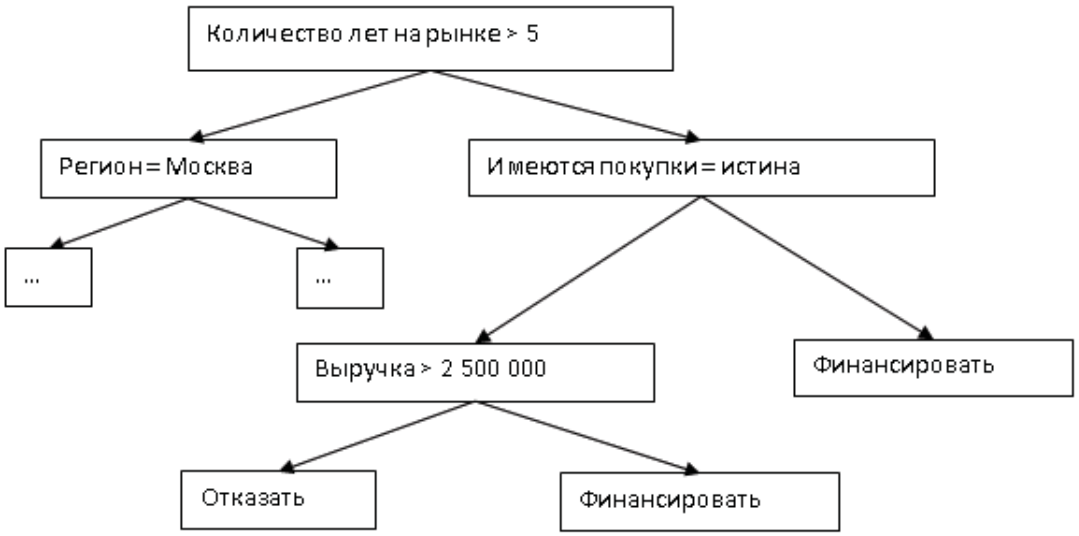
По определению Р. Квинлана [136] дерево решений представляет собой представление правил принятия решений с помощью дерева (связного ациклического графа), где каждому промежуточному узлу соответствует проверочное условие, каждому листовому узлу соответствует вариант решения.

Промежуточный узел, как правило, представляет собой простой вопрос, ответ на который даёт переход к следующему узлу, рисунок 1.

В настоящее время сфера применения деревьев решений широка, и задачи, решаемые с помощью этого аппарата можно объединить в следующие группы:

- описание данных;
- классификация;

- регрессия (установление зависимости, прогнозирование).



Источник: составлено автором.

Рисунок 1 – Пример принятия решения с помощью дерева решений

Построение дерева решений можно описать следующим алгоритмом.

Имеется обучающая выборка T, содержащая множество объектов, каждый из которых имеет m атрибутов. Один из этих атрибутов указывает на отношение объекта к одному из классов.

В работах Р. Куинлена [136] идея построения деревьев решений, впервые озвученная Хантом, формально описывается следующим образом.

Классы обозначаются множеством в соответствии с формулой (1)

$$C = \{C_1, \dots, C_k\} \quad (1)$$

тогда:

1. T содержит примеры, относящихся к одному классу C_i. В таком случае дерево решений представляет собой лист, определяющий класс C_i;

2. T не содержит примеров, т.е. является пустым множеством. Тогда это лист. Класс будет выбран из другого множества отличного, например, ассоциированного с родителем;

3. T содержит примеры из разных классов. Тогда необходимо разбить T на подмножества. Выберем признак, со значениями, отличными друг от друга, формула (2):

$$O = \{O_1, \dots, O_n\} \quad (2)$$

Разбивая T на подмножества (3)

$$T = \{T_1, \dots, T_n\} \quad (3)$$

можно увидеть, что каждое из подмножеств T_j содержит весь набор примеров, принимающих значение O_j.

Процедура рекурсивно выполняется до тех пор, пока полученное множество не станет состоять из примеров одного класса.

Примеры наиболее популярных на сегодняшний день алгоритмов, реализующих деревья решений:

- CART [67, С. 259] – один из алгоритмов построения бинарного дерева (Classification and Regression Tree). В каждую вершину графа входит не более одного ребра и выходит не более двух. Алгоритм решает задачи классификации и регрессии.
- C4.5 [135] – один из алгоритмов построения дерева с любым количеством возможных вариантов решений. Не работает с непрерывными значениями, поэтому решает только задачи классификации.

Почти все известные алгоритмы построения деревьев решений являются «жадными»: когда выбран один атрибут, по которому произведено разбиение, алгоритм не возвращается на предыдущий атрибут, чтобы улучшить разбиение. Таким образом, жадные алгоритмы не дают наилучшего разбиения.

Использование деревьев решений для прогнозирования конверсии маркетинговых коммуникаций представляется проблематичным в силу ряда причин:

1. Получение оптимального дерева решений представляет собой NP-полную задачу. Имея в наличии сотни тысяч клиентов и десятки

признаков невозможно построить оптимальное дерево и интерпретировать его смысл.

2. Методы построения деревьев решений склонны к переобучению. Это не позволяет создать устойчивую модель конверсии.

3. На практике очень часто приходится работать именно с категориальными переменными. Если категориальные переменные имеют большую вложенность уровней, то больший вес присваивается атрибутам с большим количеством уровней. Это ограничивает использование категориальных переменных.

1.2.2 Нейронные сети

Впервые концепция нейронной сети была изложена в фундаментальной работе У. Маккалока и У. Питтса «О логическом исчислении идей и нервной активности» [33]. Современные исследования этого типа моделей представлены в работах А. Котса, В. Девида, Т. Вонга [158].

Искусственная нейронная сеть – упрощённая модель мозга. Нейронная сеть является самообучающейся системой.

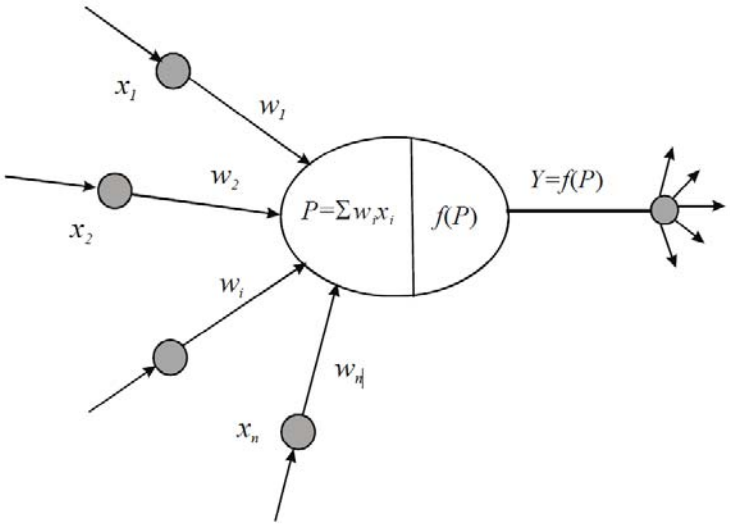
На вход первого слоя нейронной сети информация поступает из внешней среды, на вход внутренних – от других нейронов. Каждый нейрон имеет функцию преобразования сигналов (сумматор). Модель нейрона, представляет собой преобразователь информации, изображённый на рисунке 2.

Информация, полученная нейроном суммируется с весовыми коэффициентами w_i для каждого поступающего сигнала $x_i, i = 1, \dots, n$, где n – размерность пространства входных сигналов.

Потенциал нейрона вычисляется по формуле (4) [158]:

$$P = \sum_{i=1}^{n} w_i x_i \tag{4}$$

Исходящий сигнал нейрона преобразуется с помощью передаточной функции $f(P)$. Вид этой функции является важнейшим параметром нейрона. В общем случае вид функции $f(P)$ задаётся ступенчато, линейно или нелинейно, как на рисунке 3. Пороговая функция проводит сигнал только тогда, когда он превышает пороговое значение. Такой вид функции довольно прост в задании, но лишает нейронную сеть гибкости при обучении. В конечном счёте, если суммарное значение не достигает порога, выходной сигнал нейрона теряет интенсивность, что значительно влияет на результаты работы следующих нейронов в сети.



Источник: [8, С. 99].

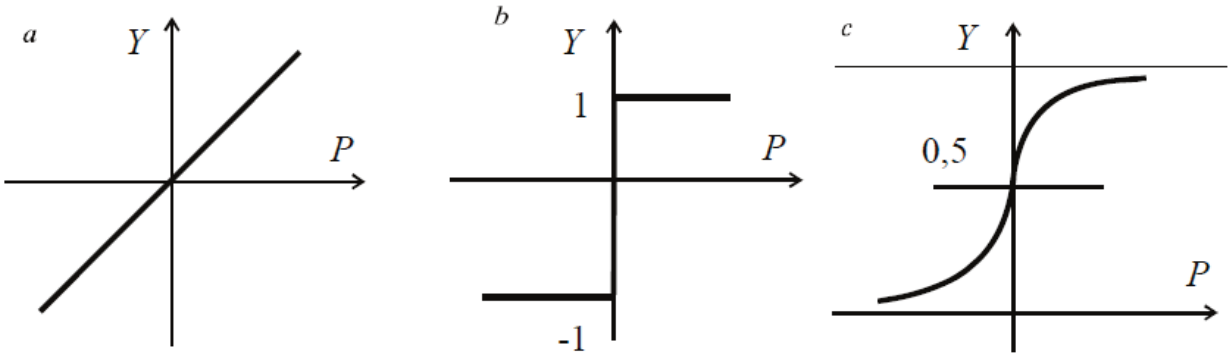
Рисунок 2 – Схема кибернетической модели нейрона

Преимуществом использования линейной функции является возможность её дифференцирования. Это позволяет уменьшить долю ошибок выходных сигналов сети. Однако, в большинстве задач используется сигмоидальная передаточная функция, представляющая собой компромисс между линейным и ступенчатым вариантом (5) [8].

$$Y = \frac{1}{1 + e^{-kP}} \tag{5}$$

Такая функция также хорошо моделирует работу биологического нейрона. Коэффициент k задаёт резкость перехода: чем он выше, тем

сигмоидальная функция ближе к пороговой, чем ниже – тем ближе она к линейной. Сигмоидальная функция дифференцируема и, в отличие от линейной и пороговой, не имеет разрывов.



Источник: [8, С. 100].

Рисунок 3 – Функции переноса искусственных нейронов: а – линейная функция; б – ступенчатая функция; в – сигмоидальная функция

В решении конкретной задачи для построения нейронной сети необходимо выбрать вид передаточной функции, её параметры, количество нейронов и количество слоёв, наличие скрытых слоёв.

Для предварительного анализа многомерных данных о клиентах можно использовать одну из разновидностей нейронных сетей – карты Кохонена. В случае смещения входных сигналов разбиение на кластеры может дать неприемлемый результат, а раскраска может оказаться монотонной. В главе III предлагаются способы коррекции таких неточностей.

При проектировании искусственных нейронных сетей неизбежно возникает ряд проблем, препятствующих их использованию в прогнозировании бинарного отклика на маркетинговые коммуникации:

- подходы к проектированию ИНС не приводят к однозначным решениям, поскольку являются эвристическими;
- моделирование на основе ИНС требует итеративной настройки внутренних элементов модели и связей между ними;

- отсутствие достаточного количества обучающих примеров представляет собой проблему для подготовки обучающей выборки;
- в ряде случаев обучение сети приводит к тупиковым ситуациям;
- невозможность применения ИНС в системах реального времени в следствие продолжительных временных затрат на обучение;
- непредсказуемость поведения обученной ИНС, увеличение риска при применении ИНС в управлении дорогостоящими техническими объектами;
- большую часть коммерческих проектов реализации нейронной сети в виде сверхбольших интегральных схем нельзя назвать легкодоступными.

1.2.3 Логистическая регрессия

Применение логистической регрессии в анализе отклика клиентов на маркетинговые предложения предпочтительнее, так как логистическая регрессия позволяет задать вероятность бинарного отклика в виде непрерывной функции. Изучению логистической регрессии посвящено множество работ [62, 64, 74, 109, 102, 113, 117, 121, 129, 132, 147, 153, 155, 150, 106, 63, 95], в частности, алгоритмам вычисления параметров [99, 73, 106, 124] и статистическим тестам [90, 81].

Любая регрессионная модель кратко записывается формулой (6) [46, С. 4]:

$$y = F(x_1, x_2, \dots, x_n) \quad (6)$$

Например, для множественной линейной регрессии эта формула будет иметь вид (7) [19, С. 6]:

$$y = a + b_1 x_1 + b_2 x_2 + \dots + b_n x_n \quad (7)$$

Формально множественную линейную регрессию можно использовать для моделирования бинарного отклика. Например, если клиент откликнулся на маркетинговую коммуникацию, значение предсказываемой переменной на

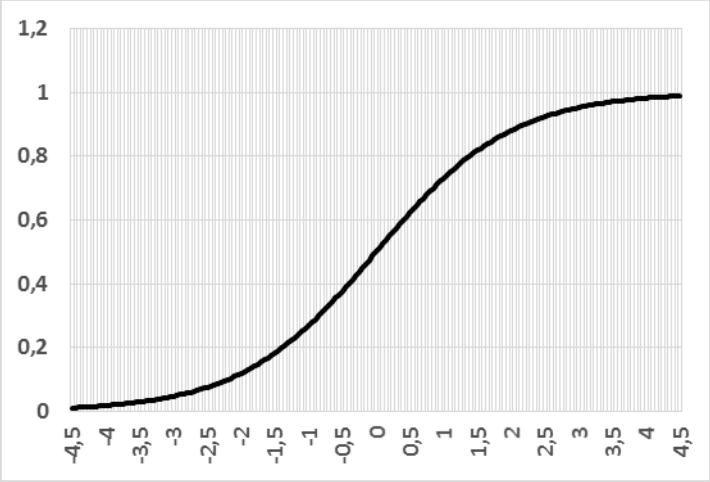
этапе обучения будет равно единице, иначе – нулю. Однако, построенная регрессионная модель (хотя бы с одним значимым коэффициентом) неизбежно будет предсказывать значения вероятности отклика, меньшие нуля и большие единицы.

Для того, чтобы этого избежать, и привести предсказываемое значение в диапазон от нуля до единицы, выполняется логит-преобразование [99, С. 1065] по формуле (8):

$$P = \frac{1}{1 + e^{-y}}, \tag{8}$$

где P – вероятность возникновения интересующего события; y – линейная комбинация значимых факторов модели.

График логистической функции представляет собой сигмоиду, представленную на рисунке 4:



Источник: составлено автором.

Рисунок 4 – Логистическая кривая

Обратное преобразование будет выглядеть следующим образом (9) [94]:

$$P' = \ln(\frac{P}{1 - P}) \tag{9}$$

Таким образом, значение линейной комбинации факторов модели y будет равен P', который может принимать значения в любом диапазоне.

Для поиска значений коэффициентов логистической регрессии часто прибегают к использованию метода максимального правдоподобия (likehood function). В основе этого метода лежит функция правдоподобия, отражающая вероятность совместного появления результатов выборки $Y_1, Y_2, \dots, Y_k$ (10) [57, 58]:

$$L(Y_1, Y_2, \dots, Y_k; \theta) = p(Y_1; \theta) \cdot \dots \cdot p(Y_k, \theta) \tag{10}$$

В качестве оценки неизвестного параметра применяется значение, максимизирующее L. Нахождение этой оценки можно упростить, если вместо L максимизировать $ln(L)$ (11):

$$L^*(Y, \theta) = \ln\big(L(Y, \theta)\big) \to max \tag{11}$$

В случае бинарной переменной через P_i обозначается вероятность появления единицы: $P_i = Prob(Y_i = 1)$. Эта вероятность зависит от $X_i W$, где W – вектор коэффициентов регрессии, X_i – вектор-строка в матрице регрессоров (12, 13) [106, С. 32]:

$$P_i = F(X_i W) \tag{12}$$

$$F(z) = \frac{1}{1 + e^{-z}} \tag{13}$$

Тогда функция правдоподобия будет иметь вид (14):

$$\begin{aligned} L^* = &\sum_{i \in I1} lnP_i(W) \\ &+ \sum_{i \in I0} \ln\big(1 - P_i(W)\big) \\ &= \sum_{i=1}^{k} [Y_i lnP_i(W) + (1 - Y_i)\ln(1 - P_i(W))], \end{aligned} \tag{14}$$

где $I0$, $I1$ – наблюдения, для которых $Y_i = 0$ и $Y_i = 1$ соответственно.

Градиент g и гессиан H функции правдоподобия будут соответственно равны (15, 16):

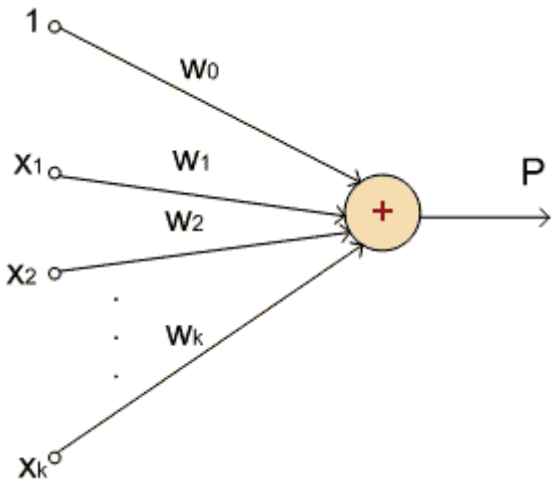
$$g = \sum_i (Y_i - P_i) X_i \tag{15}$$

$$H = -\sum_i P_i(1 - P_i)X_i^T X_i \leq 0 \quad (16)$$

Для того, чтобы найти максимум, часто прибегают к использованию метода Ньютона (17):

$$W_{t+1} = W_t - \left(H(W_t)\right)^{-1} g_t(W_t) = W_t - \Delta W_t \quad (17)$$

Логистическую регрессию также представляют в виде однослойной нейронной сети, в которой функция активации имеет сигмоидальный вид, рисунок 5.



Источник: [40].

Рисунок 5 – Логистическая регрессия в виде нейронной сети

1.2.4 ROC-анализ

Для анализа качества бинарной классификации применяется ROC-анализ. Исследования, посвящённые этой теме, представлены в работах [59, 70, 89, 96, 99, 104, 116, 145, 149, 158, 159]. При бинарной классификации могут возникать ошибки первого и второго рода. В таблице 2 приведены варианты классификации, где:

- TP (True Positives) – истинно-положительная классификация;
- TN (True Negatives) – истинно-отрицательная классификация;
- FN (False Negatives) – ложно-отрицательная классификация;

- FP (False Positives) – ложно-положительная классификация.

Таблица 2 – Матрица сопряжённости

Исходы		**Фактически**	
		Положительно	Отрицательно
Модель	Положительно	TP	FP
	Отрицательно	FN	TN

Источник [40].

При анализе часто пользуются относительными показателями

True Positives Rate (TPR) – доля истинно положительных классификаций (18):

$$TPR = \frac{TP}{TP + FN} \cdot 100\% \quad (18)$$

False Positives Rate (FPR) – доля ложно положительных классификаций примеров (19):

$$FPR = \frac{FP}{TN + FP} \cdot 100\% \quad (19)$$

Вероятность положительного исхода при положительной классификации (precision или positive predictive value) PPV (20):

$$PPV = \frac{TP}{TP + FP} \quad (20)$$

Чувствительность (Sensitivity) – доля истинно положительно классифицированных случаев (21):

$$Se = TPR = \frac{TP}{TP + FN} \cdot 100\% \quad (21)$$

Специфичность (Specificity) – подразумевает долю истинно отрицательных классификаций отрицательных примеров (22):

$$Sp = \frac{TN}{TN + FP} \cdot 100\% \quad (22)$$

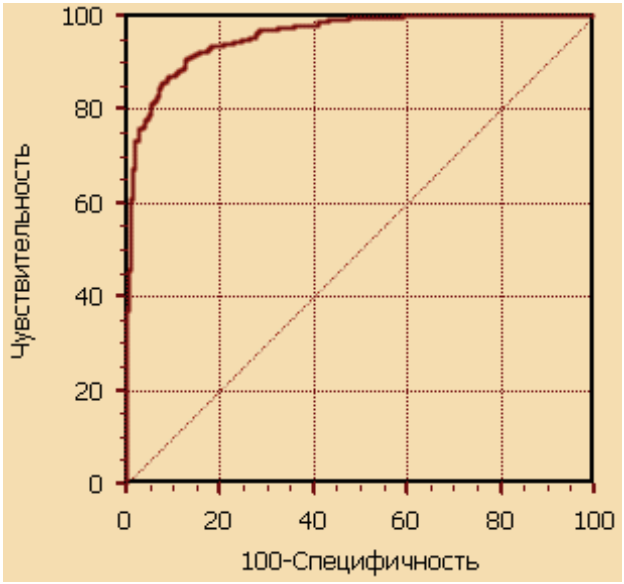
График ROC-кривой представлен на рисунке 6.

AUC (area under curve) – площадь под графиком ROC-кривой. Чем она выше, тем качество классификации лучше, рисунок 7.

AUC (Area Under Curve) Можно вычислить следующим образом (23):

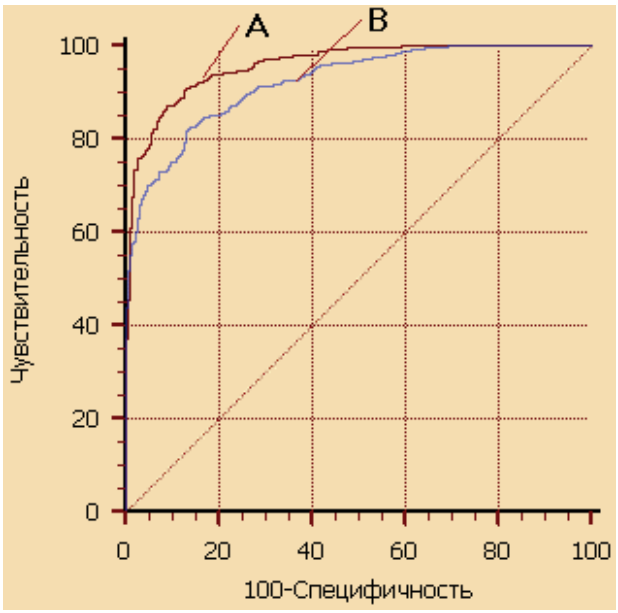
$$AUC = \int f(x)dx = \sum_i \frac{X_{i+1} + X_i}{2} \cdot (Y_{i+1} - Y_i) \quad (23)$$

На практике применяются разные формальные критерии для выбора порога отсечения. В случае с анализом маркетинговых коммуникаций задаётся минимальная вероятность отклика (PPV) и максимизируется чувствительность.



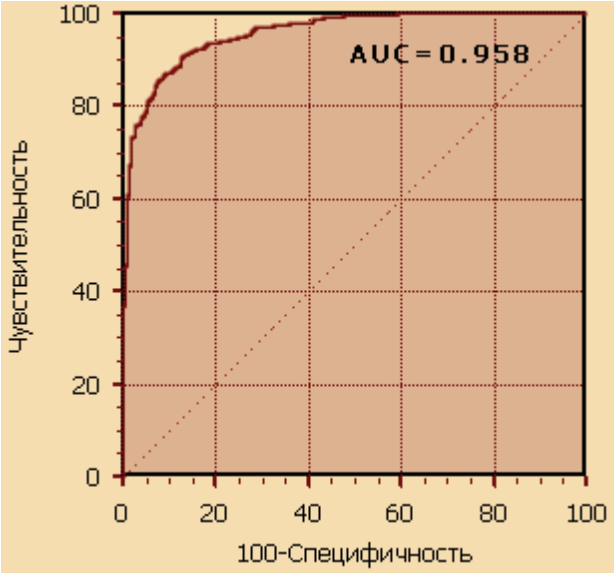
Источник [40].

Рисунок 6 – Кривая ROC



Источник: [40].

Рисунок 7 – Сравнение ROC-кривых



Источник: [40].

Рисунок 8 – Площадь под ROC-кривой

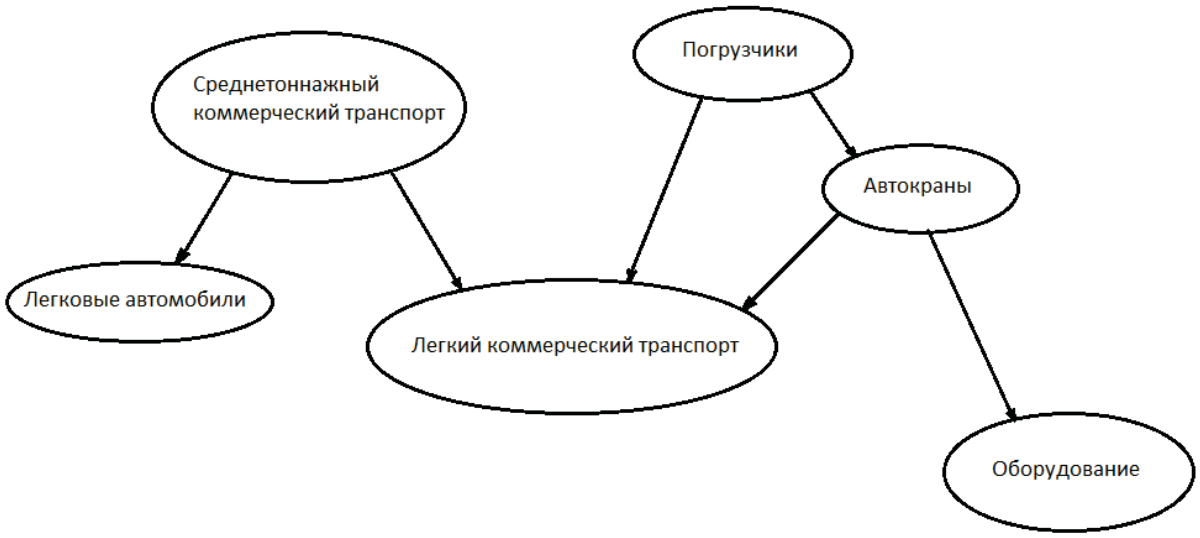
В силу относительной простоты по сравнению с другими типами моделей логистическая регрессия не всегда даёт результат желаемой точности. Однако, по этой же причине её результаты хорошо интерпретируемы и имеют высокую степень надёжности. Значение, прогнозируемое логистической регрессией является непрерывным, таким образом, этот инструмент представляется наиболее подходящим для прогнозирования вероятности бинарного отклика в маркетинговых коммуникациях.

1.2.5 Ассоциативные правила

Ассоциативные правила разработаны в компании IBM [54] в исследовательском центре IBM Almaden. Этот инструмент позволяет искать закономерности между событиями, рисунок 9. Утверждение «покупатель, приобретающий грузовики, приобретёт и погрузчики с вероятностью 75%» будет представлять собой пример такого правила.

Формально постановка задачи имеет следующий вид. Набор покупательских транзакций зафиксирован в базе данных. Каждая такая

транзакция представляет собой набор товаров, которые были куплены покупателем за один раз. Ещё такие транзакции называют «рыночной корзиной».



Источник: составлено автором.

Рисунок 9 – Пример ассоциативных правил

Множество $I = \{i_1, \dots, i_n\}$ – есть множество товаров, D – есть множество транзакций, в котором каждая транзакция T – представляет собой набор элементов из множества I, где $T \subseteq I$. Каждая транзакция представляется в виде бинарного вектора, где $t[k] = 1$, в случае, если i_k-й элемент имеется в транзакции, в противном случае $t[k] = 0$. Говорят, что транзакция содержит X, где X – это набор элементов из множества I, $X \subset T$. Импликация $X \Longrightarrow Y$, где $X \subset I$, $Y \subset I$ и $X \cap Y = \emptyset$ называется ассоциативным правилом. Для правила $X \Longrightarrow Y$ существует поддержка s, когда s% транзакций из множества D, имеют $X \cup Y$, $supp(X \Longrightarrow Y) = supp(X \cup Y)$. Вероятность того, что из X следует Y называют *достоверностью правила*. Правило $X \Longrightarrow Y$ с достоверностью (confidence) c является справедливым, когда c% транзакций из множества D, которые содержат X, содержат и Y (24):

$$conf(X \Longrightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \quad (24)$$

Алгоритм Apriori [53] – это один из первых алгоритмов, который был способен эффективно решать подобный класс задач. В последнее время кроме него были разработаны и другие алгоритмы: DHP[141], Partition[70], DIC[68].

Обобщённые ассоциативные правила позволяют связывать различные сущности, например, группы с группами, или отдельные элементы с группами.

Импликация $X \Longrightarrow Y$, в которой $X \subset I$, $Y \subset I$ и $X \cap Y = \emptyset$ где ни один из элементов множества Y, не является предком хотя бы одного элемента множества X называется обобщённым ассоциативным правилом. Для обобщённых ассоциативных правил поддержка и достоверность вычисляются аналогичным образом.

Дополнение информации об иерархической группировке может дать такие преимущества, как:

- выявление ассоциативных правил между уровнями иерархии, а не только между какими-либо отдельными элементами;
- группа может удовлетворять порогу minsupport, хотя отдельные элементы имеют риск получить недостаточную поддержку.

Любой из вышеназыванных алгоритмов подойдёт для нахождения таких правил. Каждая транзакция дополняется всеми предками каждого входящего в неё элемента. Тем не менее, применение этих алгоритмов неизбежно приводит к ряду проблем:

1. Элементы на нижних уровнях иерархии устремляются к значительно меньшим значениям поддержки по сравнению с элементами на верхних уровнях.

2. Размерность входного пространства напрямую зависит от количества групп, в которые объединены транзакции. Это приводит к созданию большего количества правил и неизбежно усложняет задачу.

3. Возможно появление избыточных правил, таких как «Автомобили» - «Грузовики». Даже при 100% достоверности такое правило будет иметь нулевую практическую ценность. Таким образом возникает необходимость использования специальных операторов, удаляющих подобные избыточные правила.

В работе [54] были предприняты попытки решения указанных проблем. Предлагаемые в ней методы имеют лучшее быстродействие, чем Apriori и устраняют такие проблемы.

Также элементы можно группировать по другим характеристикам, таким как цена, бренд, сегмент и пр.

Численные ассоциативные правила позволяют комплексно учесть не только факт покупки товаров, но и характеристики покупателя, количество купленного товара с учётом разных типов данных. Алгоритмы поиска численных ассоциативных правил предлагаются в работе [56].

Применение ассоциативных правил в анализе маркетинговых коммуникаций позволяет не только выяснить взаимосвязи между покупаемыми продуктами, но и применить эти знания в построении модели отклика на маркетинговые коммуникации, в частности, устранить негативное влияние корреляции между зависимыми признаками.

1.2.6 RFM-анализ

RFM-анализ является популярным методом исследования поведенческих характеристик клиента. Одни из последних разработок в этом направлении представлены в работе К. Ли, П. Фадера и Б. Харди [118, 85].

Постановка задачи формулируется следующим образом: в заданный момент времени имея данные клиентской истории необходимо определить: стоит ли подготовить маркетинговое предложение для рассматриваемого клиента «А».

Говоря более формальным языком, набор данных в поставленной задаче D записывается в таблицу из N элементов (записей), характеризующихся l различными параметрами $h_1, \dots, h_l$. Параметры могут быть как числовыми, так и категориальными. Запись (строка) $X_i = (h_1, \dots, h_l)$ отражает историю действий, совершаемых клиентом X_i в момент создания таблицы D. Все N элементов заведомо определены в одну из двух групп (классов) C_1 и C_2. В практических приложениях только одна из этих двух групп является целевой и, как правило, также является миноритарной. Эту группу ещё называют позитивной группой, а наблюдения из рассматриваемой группы – позитивными наблюдениями. Вторая группа (класс) называется негативной. На практике позитивную группу обозначают единицей, а негативную – нулём.

Задача сегментации сводится к построению классификатора, который позволяет:

1. Оценить вероятность принадлежности наблюдения позитивной группе $P_i = F(h_{1i}, \dots, h_{li})$. Эта вероятность также называется скором (от англ. score – очки). Модель классификатора, таким образом, называется скоринговой моделью.

2. Сгруппировать схожих по вероятности отклика клиентов в $k > 2$ сегментов.

В большинстве работ, посвящённых задаче целевого выбора, авторы ограничиваются первым пунктом приведённой постановки задачи.

Многие исследователи в сфере директ-маркетинга приходят к одному выводу: факторы поведенческой истории клиента оказывают гораздо более сильное влияние на принадлежность наблюдений позитивной группе, чем результаты анкетирования клиента [154]. Это привело к тому, что большинство моделей сегментации используют универсальный, но ограниченный набор переменных, известных как RFM по первым буквам названий переменных.

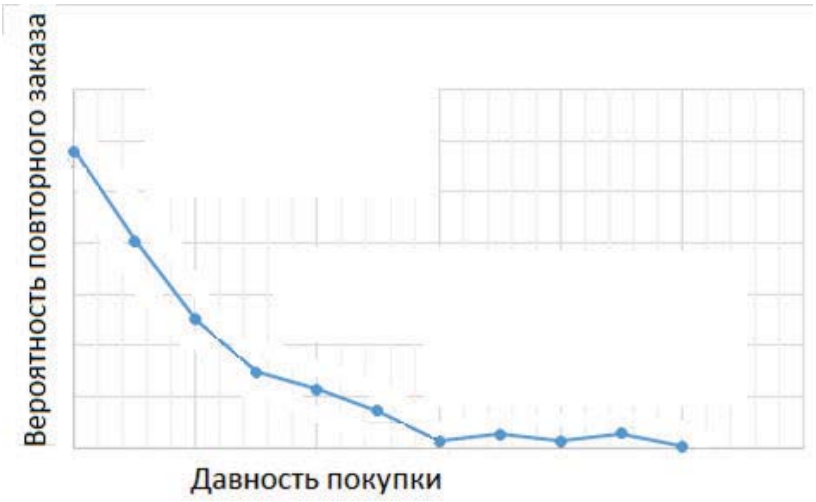
- Recency (R) – время, прошедшее с момента последнего заказа, рассчитанное в сутках;
- Frequency (F) – суммарное количество заказов клиента за всю историю покупок;
- Monetary (M) – суммарное количество денег, затраченных клиентов на покупки.

RFM-сегментация используется для прогнозирования принадлежности позитивному классу ввиду того, что история покупательского поведения часто является надёжным ориентиром, указывающим на будущее покупательское поведение [138].

Ниже описываются типовые зависимости вероятности повторного заказа (принадлежности позитивному классу) от RFM переменных, подтверждаемые большим количеством экспериментальных данных.

R-постулат. Вероятность очередной покупки отрицательно зависит от давности последней покупки.

Типичный вид кривой представлен на рисунке 10.



Источник: составлено автором на основе данных [42].

Рисунок 10 – Зависимость вероятности повторного заказа от давности покупки

F-постулат. Вероятность покупки положительно зависит от частоты покупок: чем чаще клиент делает покупки, тем выше вероятность нового заказа.

Типичный вид кривой представлен на рисунке 11.

Источник: составлено автором на основе данных [42].

Рисунок 11 – Зависимость вероятности повторного заказа от количества сделанных заказов почти линейная с насыщением в области больших значений

M-постулат. Вероятность покупки положительно зависит от суммарно затраченных клиентом денег: чем больше потрачено денег, тем выше вероятность будущего заказа.

Типичный вид кривой показан на рисунке 12.

Предсказательная мощность перечисленной тройки переменных ранжируется аналогично последовательности букв в названии RFM: Recency является наилучшим предиктором, вторым идет Frequency и последним по силе является Monetary [143, С. 113].

Имеет смысл отметить что, переменные RFM коррелируют между собой. Как уже упоминалось выше, частота при наступлении привычки и все уменьшающийся промежуток времени между повторными, привычными

действиями объясняет эмпирический постулат Recency. Соответственно, чем меньше Recency, тем больше Frequency, т.е. R и F коррелируют между собой отрицательно. Так же очевидно, что факторы M и F имеют прямую зависимость: чем больше покупок совершает клиент, тем большее количество средств он на них тратит. Соответственно M и F коррелируют положительно. Аналогично паре F и R, корреляция между R и M отрицательна.

Источник: составлено автором на основе данных [42].

Рисунок 12 – Зависимость вероятности повторного заказа от количества денег, потраченных на все заказы имеет квазилинейный вид с насыщением в области больших значений

В большинстве приложений, результатом построенной модели является вещественное число p, принимающее значение в интервале [0, 1], и трактующееся как вероятность принадлежности записи $X_i = (h_1, \dots, h_l)$ позитивной группе. При решении задачи скоринга применяют такие методы прикладной статистики, как линейная и нелинейная регрессия [5, 3, 9, 10, 137], нейронные сети [18, 22, 154], деревья решений [10, 14], дискриминантный анализ [1, 10, 14], ассоциативные правила [118, 54, 53, 137] и др.

Сравнивая различные классические подходы методом кросс-проверки и методом построения карт выигрышей [137] на идентичном наборе данных с RFM предикторами выявляется сопоставимое качество построенных классификаторов. Перечисленные методы дают достаточную эффективность и применяются в различных директ-маркетинговых кампаниях при решении задачи целевого выбора.

Помимо инструментария прикладной статистики при решении задачи сегментации применяют прочие методы, основанные на RFM-переменных. Наиболее простым и широко используемым сегодня методом сегментации клиентов является метод RFM-кодирования, или просто RFM-сегментация [143, 125]. Суть RFM-кодирования состоит в том, что значения переменных (R, F и M) разделяются на 5 интервалов по квинтилям распределения клиентов, или по квинтилям распределения значений переменных. Полученные группы кодируются значениями от 1 до 5 в соответствии с постулатами о поведении клиентов: единица присваивается группе с наименьшей вероятностью отклика, пятёрка – с наибольшей.

Существующие методы сегментации клиентской базы данных имеют общий недостаток – отсутствие фиксирования границ отсечения по вероятности отклика. Несмотря на то, что методы позволяют решать поставленную задачу этот недостаток является существенным.

Границы отсечения сегментов изменяются в зависимости от распределения сегментов по RFM-переменным. Таким образом, при повторной сегментации может случиться, что сегмент k по R будут образовывать клиенты с $R < 20$, тогда как при предыдущей сегментации этот же сегмент образовывали клиенты с $R < 15$.

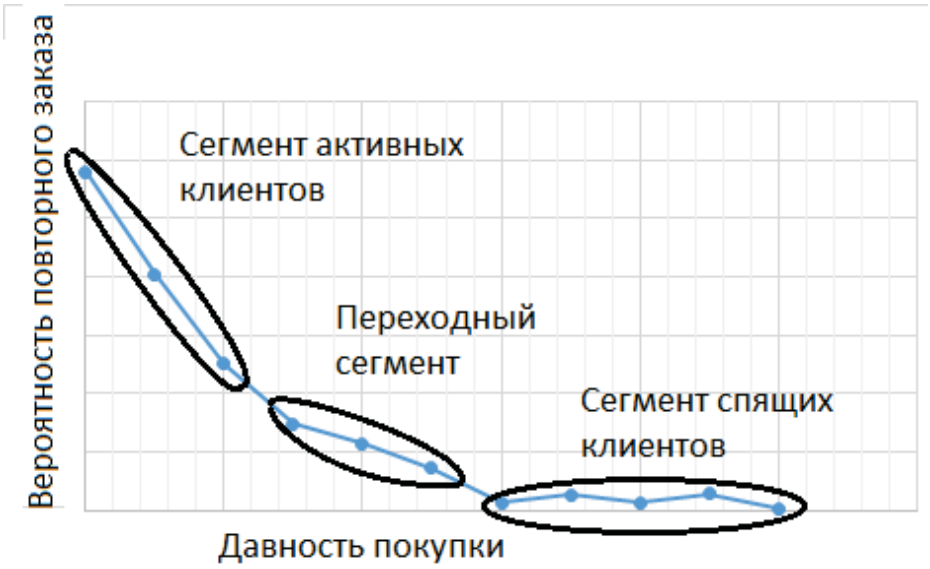
Таким образом из-за отсутствия фиксации границ методы сегментации не позволяют:

- в рамках сегментов делать выборочные тесты новых маркетинговых концепций, находить оценки вероятности отклика на них для дальнейшего применения результатов на остальной части сегмента в силу

того, что изменение границ сегментов приводит к ненадёжности оценок статистических тестов [137],

- отслеживать динамику изменения сегментов клиентской базы данных, так как классы не являются идентичными на протяжении всех моментов времени.

Фактор давности покупки сам по себе является достаточным для сегментации клиентов [125]. Он описывает наиболее значимые этапы жизненного цикла клиента. Иначе говоря, кривая фактора Recency графически описывает жизненный цикла клиента, рисунок 13.



Источник: составлено автором на основе данных [42].

Рисунок 13 – Кривая давности последнего заказа R

Зона активных клиентов – это область больших значений вероятности отклика для малых значений Recency. Клиенты в этой зоне находятся на пике своей активности и приносят компании максимальную прибыль. Зона спящих клиентов – это область малых значений вероятности отклика и относительно больших значений Recency. Клиенты в этой зоне находятся в стадии «доживания» своей истории взаимодействия с компанией. Использование этих клиентов на постоянной основе довольно рискованно.

После получения трёх сегментов активных, спящих и переходных клиентов разбиение проверяется на статистическую значимость. Обычно для этой цели используется биномиальный критерий. Таким образом, на значимость проверяются различия в вероятностях отклика клиентов, принадлежащих двум парам сегментов:

- сегмента активных клиентов и сегмента переходных клиентов. Порог $p < 0.05$;
- сегмент спящих клиентов и сегмент переходных клиентов. Порог $p < 0.05$.

Различие между сегментами активных и спящих клиентов следует автоматически при соблюдении перечисленных условий.

В случае сегментации по фактору Frequency последним двум сегментам уделяется особенное внимание, поскольку сегмент активных клиентов по своей сущности не требует дополнительных группировок. Таким образом, сегментация по Frequency ставит перед собой две цели:

- уменьшение неопределённости переходного сегмента;
- нахождение спящих клиентов, которые наиболее склонны к получению маркетингового предложения.

Для сегментации по фактору Frequency разработаны два правила: правило отсечения одного заказа и правило группировки по процентилям.

В большинстве клиентских баз данных содержится большая доля клиентов, которые совершили всего одну покупку и перешли в сегмент спящих. По разным оценкам доля этого сегмента составляет от 30% до 60% [125]. Имеет смысл упомянуть, что последующее поведение клиента, который сделал всего один заказ сильно отличается от поведения клиентов, совершивших большее количество покупок. По мнению маркетологов это можно объяснить тем, что клиент, делая всего один заказ, пробует предлагаемую продукцию, а делая второй заказ подтверждает свою приверженность к бренду, и, таким образом, становится постоянным клиентом. От 30% до 60% клиентов делают пробную покупку и более не

проявляют желания взаимодействовать с компанией. Очевидно, что такую многочисленную группу клиентов, которая имеет схожую историю поведения, а значит будет иметь схожее поведение в будущем, имеет смысл анализировать отдельно. Для учёта такого немаловажного факта используют правило отсечения одного заказа при сегментации по фактору Frequency.

Правило отсечения одного заказа по фактору Frequency подразумевает разбиение сегментов спящих и переходных клиентов на две группы:

- клиенты с одной покупкой;
- клиенты с двумя и более покупками (2+).

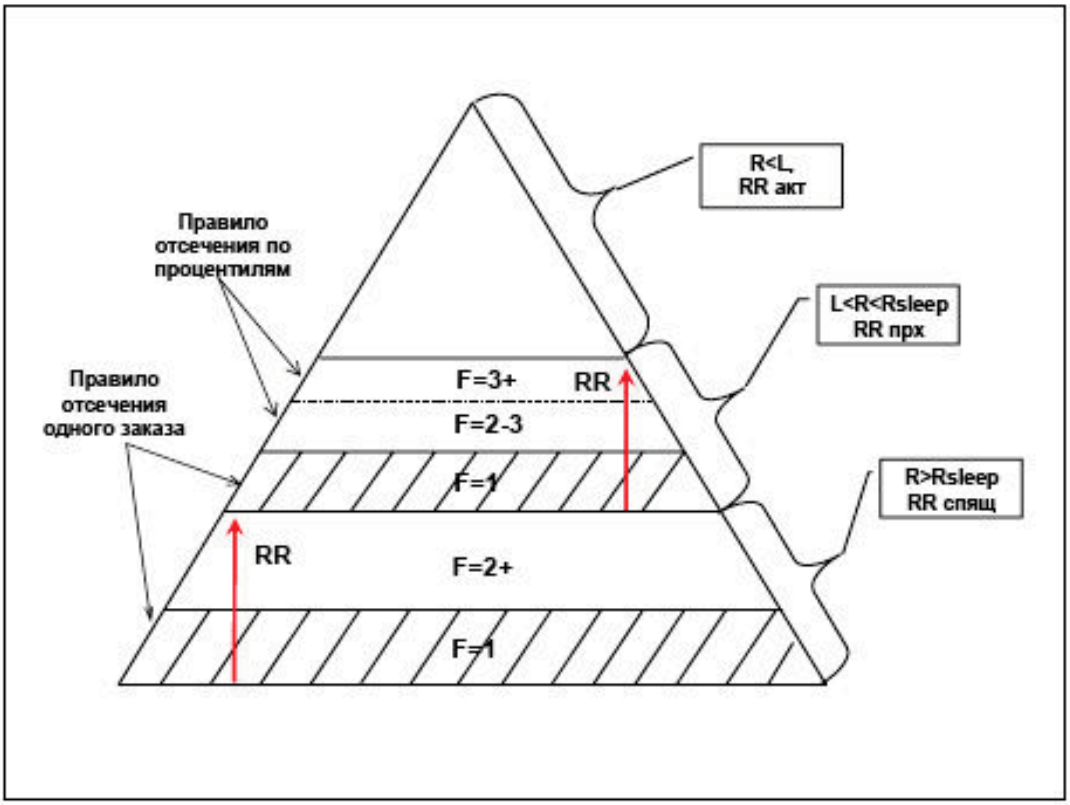
После проведения отсечения группы клиентов с одной покупкой вычисляют долю клиентов с двумя и более покупками. Если второй сегмент оказывается слишком большим, его разбивают ещё на две группы. F-постулат клиентского поведения позволяет утверждать, что дополнительное разбиение по F позволит выявить наиболее интересный с точки зрения маркетинговых коммуникаций сегмент, а также сократить неопределённость, обусловленную переходным сегментом. В случае с сегментом спящих клиентов дополнительная группировка по фактору Frequency позволит выделить группу клиентов для восстановления.

Алгоритм проверки разбиения на значимость представляется следующим образом:

- для всех $i \in \{1, \dots, k-1\}$ пока не выполнится $pB_i(S_i, S_i+1) < 0{,}05$ выполнять: Если $pB_i(S_i, S_i+1)$, то $S_i = S_i \cup S_i + 1$, где $pB_i(S_i, S_i+1)$ – есть p-уровень значимости для биномиального критерия, который применяется последовательно к двум сегментам $S_i, S_i + 1$.

На рисунке 14 представлена сегментация по фактору Frequency. В начале применяют правило отсечения одного заказа, потом для второго сегмента применяют правило отсечения по процентилям. В текущем примере второе правило затронуло только переходный сегмент в области, где значения Frequency переходят от двух к трём. Что касается сегмента спящих

клиентов, то для них результаты группировки по процентилям получаются незначимыми. Внутри сегментов «направления» возрастания вероятности отклика указаны стрелками (в соответствии с F-постулатом).



Источник: [42].

Рисунок 14 – Сегментация по фактору Frequency

За отчётный период можно произвести расчёт следующих показателей за счёт сегментации:

- Коэффициент трансформации (transformation rate, TR) – процент клиентов из сегмента, которые проявили активность;
- Ords per customer (OPC) – количество заказов, приходящихся на одного клиента из сегмента;
- Ords per active customer (OPaC) – количество заказов, приходящихся на одного активного клиента из сегмента;
- Turnover (T) – размер оборота, приходящегося на сегмент;

- Turnover per customer (TPC) – размер оборота, приходящегося на одного клиента из сегмента;
- Turnover per active customer (TPaC) – оборот на одного активного клиента;
- AOV – средний чек по всему сегменту.

Имея в наличии дополнительные поведенческие и финансовые показатели можно выяснить, какой из сегментов приносит наибольшую выручку, какое количество покупок сделано клиентами из одного сегмента и какова средняя стоимость покупки. Как правило для сегмента активных клиентов выполняется правило Парето: 20% активных клиентов приносят 80% выручки.

RFM-анализ, не смотря на широкую популярность не всегда разумно использовать в его изначальном виде. В частности, в решении задачи оптимизации маркетинговых коммуникаций и привлечения клиентов некоторые факторы выгоднее заменить, а также дополнить модель другими найденными значимыми показателями. Таким образом, дополнение общих принципов RFM-анализа знаниями из предметной области и их реализация с помощью регрессионного анализа (в частности логистической регрессии) является сильным подходом к прогнозированию вероятности отклика клиентов и «выжиманию» из клиентской базы максимальной прибыли.

1.2.7 Customer Lifetime Value

CLV (customer lifetime value) – параметр, отражающий прибыль компании за счёт одного клиента на протяжении всего времени взаимодействия с ним. Методы и проблематика расчёта CLV рассматриваются в работах [85, 141, 134, 138, 61, 74, 115, 109].

Перед приведением формулы расчёта CLV, следует упомянуть:

- показатель следует вычислять для каждого канала привлечения в отдельности. Не смотря на то, что общее значение может быть удовлетворительным, канал привлечения может оставаться убыточным;
- некорректно применять показатель для клиентов, совершивших всего одну покупку;
- необходимо сравнивать показатели за аналогичный период;
- если сравниваются два канала, то сравнивается только конечный результат, а не расчётные единицы. Привлечение клиентов производится по различным каналам. Сохранение клиента может стоить дороже, но при этом не означать плохую работу с ним.

Итак, CLV - это доход, полученный от клиента, за вычетом затрат на его привлечение и удержание.

CLV рассчитывают для каждого цикла покупок.

Для примера возьмём следующие исходные данные:

1-я покупка – 2000 клиентов, 15000 р. средний чек, стоимость привлечения – 3000 р., себестоимость – 8500 р.;

2-я покупка – 900 клиентов (предположим, что вторую покупку совершают 45% клиентов), средний чек – 4000 р. (покупаются только расходные материалы), стоимость удержания – 600 р., себестоимость расходных материалов – 500 р.

Итого CLV для 1-ой покупки (25):

$$\begin{aligned} \mathrm{CLV} &= 2000 * 15000\ \text{р.} - 2000 * 8500\ \text{р.} - 2000 * 3000\ \text{р.} \\ &= 300\,000\ \text{р.} - 170\,000\ \text{р.} - 60\,000\ \text{р.} = 7\,000\,000\ \text{р.} \end{aligned} \quad (25)$$

В пересчёте на одного клиента для первой покупки (26):

$$\mathrm{CLV} = 7\,000\,000\ \text{р.}/2000 = 3500\ \text{р.} \quad (26)$$

Итого CLV для 2-ой покупки (27):

$$\begin{aligned} \mathrm{CLV} &= 900 * 4000\ \text{р.} - 900 * 500\ \text{р.} - 900 * 600\ \text{р.} \\ &= 3\,600\,000\ \text{р.} - 400\,500\ \text{р.} - 500\,300\ \text{р.} = 2\,610\,000\ \text{р.} \end{aligned} \quad (27)$$

В пересчёте на одного клиента для второй покупки (28):

$$\text{CLV} = 2\,600\,200\ \text{р.}/900 = 2900\ \text{р.} \quad (28)$$

И т.д. в зависимости от количества покупок в течение жизненного цикла клиента.

Данный пример иллюстрирует, что покупка основного продукта выгоднее для компании, чем покупка расходных материалов, которые приобрели только в 45% случаях. Тем не менее, такое резкое падение количества клиентов на второй покупке имеет место в любом бизнесе. Как правило, далее сокращение числа клиентов идёт плавнее, и постепенно достигает своего минимума: считается, что до порядка 5% от первоначального значения. Поэтому, дополнительно необходимо брать в расчёт прирост клиентской базы и её жизненный цикл.

Применение:

- один из основных показателей качества клиентской базы;
- используется в создании программ лояльности, поскольку выражает стоимость клиента и приносимую им прибыль для компании в общем и по отдельным покупкам. Вычислив CLV можно найти "золотого" клиента;
- используется в медиапланировании: рассчитав, сколько денег в среднем приносит клиент, можно оценить рекламный бюджет для всех каналов привлечения;
- даёт оценку качества работы с клиентом: если работа с клиентом организована правильно, расходы на его привлечение и обслуживание будут уменьшаться, а прибыль – увеличиваться;
- сегментировать покупателей можно не только по среднему чеку и каналу привлечения, следует добавлять также социально-демографическую информацию, территориальные и другие важные для бизнеса показатели. Таким образом можно лучше выявить прибыльные сегменты.

Вычисление CLV позволяет оценить качество маркетинговых коммуникаций с точки зрения «стоимости ошибки»: если клиент ушёл по

причине слишком частых звонков или сообщений, этот метод позволяет количественно оценить стоимость такой потери.

1.3 Инструментальные средства обработки и анализа данных

Для анализа данных в исследовании использовались программные продукты, описанные ниже.

R – это язык программирования, разработанный для статистической обработки числовой информации, а также для работы с графической информацией. R разработан в рамках проекта GNU и является свободно распространяемой вычислительной программной средой с открытым исходным кодом. В R поддерживается большой набор статистических и численных методов, этот продукт имеет хорошую расширяемость посредством подключения дополнительных пакетов, создаваемых независимыми разработчиками по всему миру. Пакеты – это библиотечные файлы с реализацией специфических функций для использования в специальных областях. В базовую комплектацию R включён набор основных пакетов. По состоянию на 2013 год разработано более 4000 пакетов.

Deductor – аналитическая платформа, в которой реализованы технологии позволяющие решать весь спектр задач полноценного анализа данных от консолидации и отчётности до прогнозирования и оптимизации [79].

В программе реализован следующий функционал:

- совмещение данных из разнотипных источников;
- очистка, систематизация, обогащение данных;
- отчетность, визуализация, OLAP-анализ, расчет KPI;
- моделирование, прогнозирование, оптимизация;
- адаптация моделей и самообучение на новых данных.

Deductor может быть использован в любом бизнесе, где используются большие объёмы данных, таблица 3.

Таблица 3 – Задачи, решаемые с помощью Deductor

Отрасль	**Типовые задачи**
Банки и страховые компании	Скоринг, риск-менеджмент, противодействие мошенничеству, противодействие отмыванию денег, отчётность
Розничные сети	Стимулирование продаж, консолидация данных, оценка лояльности, прогнозирование спроса
Дистрибьюторские фирмы	Оптимизация запасов, клиентская аналитика, контроль за ассортиментом, отчётность
Государственные учреждения	Планирование бюджета, контроль, расчёт KPI
Телекоммуникации	Наблюдение за клиентской базой, снижение оттока, адресные сообщения, увеличение доходности
Промышленность	Контроль качества, оптимизация работы оборудования, диагностика

Источник: составлено автором на основе данных [79].

VidaExpert – программный продукт, разработанный доктором физико-математических наук, Андреем Зиновьевым (Институт вычислительного моделирования, г. Красноярск).

VidaExpert позволяет:

- выполнять анализ методом взвешенных главных компонент;
- выполнять анализ методом упругих карт – аппроксимация данных нелинейными многообразиями различных топологий;
- выполнять k-means кластеризацию;
- выполнять иерархическую кластеризацию;
- выполнять дискриминантный анализ;
- строить линейную регрессию;
- строить деревья решений.

1.4 Актуальные проблемы обеспечения и поддержки качества данных, предназначенных для автоматической обработки

Чтобы обеспечить доступ к точным и согласованным данным, необходимо консолидировать различные представления данных и исключить дублирующуюся информацию.

Для обеспечения и всесторонней поддержки преобразования данных используют хранилища данных. Хранилища данных предназначены для постоянного обновления больших объёмов данных из различных источников. Они используются для принятия решений, таким образом, чтобы предотвратить получение некорректных выводов на основе анализа некорректных данных, нужно проводить регулярные корректировки. Дублирующаяся или отсутствующая информация может привести к получению ложной или неадекватной статистики.

Исследователи информационных систем выделяют основные цели использования хранилищ данных [64, С. 151], таблица 4.

Таблица 4 – Использование хранилищ данных

Типы работ	**Доля компаний-пользователей**	**Количество запросов в день**
Поддержка принятия решений	0,703	3,45
Мониторинг	0,594	2,82
Планирование	0,437	0,73
Прогнозирование	0,312	0,63
Администрирование	0,297	2,18
Учёт	0,297	1,47
Управление ресурсами	0,203	0,33
Кадровые службы	0,078	0,02
Другое	0,187	0,98

Источник: составлено автором на основе данных [64, С.151].

В этих работах также разработаны методы и приведены оценки качества данных в хранилищах данных компаний [64, С.152], таблица 5.

Таблица 5 – Показатели качества данных

Показатель	**Среднее значение**
Ценность (финансовая)	4,9/7
Уровень детализации	4,9/7
Целостность	4,1/7
Точность, достоверность	4,6/7

Источник: составлено автором на основе данных [64, С. 152].

Таким образом, очистка и структуризация данных является одним из ключевых направлений повышения их качества.

Методы очистки данных должны удовлетворять набору требований:

1. Они должны идентифицировать и устранять все основные несоответствия и ошибки, как при обработке отдельных источников данных, так и при их интеграции.

2. Эти методы должны быть поддерживаемыми инструментальными средствами для сокращения объёма ручных проверок и программирования и быть гибкими в работе с дополнительными источниками.

3. Очистка данных не должна выполняться отдельно от метаданных, связанных со схемой их преобразования.

4. Необходима поддержка инфраструктуры технологического процесса для обеспечения эффективного и надёжного выполнения всех этапов преобразования данных из всего множества источников.

5. Сообщество исследователей уделяло достаточно малое внимание очистке данных. Ряд исследовательских групп ведут проекты по решению общих проблем, прямо или косвенно связанных с очисткой данных, например, data mining.

ГЛАВА 2
РАЗРАБОТКА АЛГОРИТМОВ ПРЕОБРАЗОВАНИЯ СЛБОСТРУКТУРИРОВАННЫХ ДАННЫХ

2.1 Преобразование метрики Дамерау-Левенштейна для вычленения текстовых данных заданного типа и поиска дублирующихся записей

В прикладных задачах нечёткого сравнения часто требуется не только сравнить сущности одного класса, но и выявить их. Предлагается расширенное определение метрики Дамерау-Левенштейна, на основе которого разработан алгоритм нечёткого сравнения не только конкретных строковых последовательностей, но и некоторых их классов [39].

Извлечение данных из текста, набранного человеком является одной из наиболее актуальных проблем на сегодня. При наличии в клиентской базе большого массива данных такого рода, а также непрерывного поступления таких данных из различных внешних источников, ручная обработка такой информации невозможна.

Исследователями предложено большое количество технических средств и методов обработки текстовой информации. Вот лишь некоторые из них:

Методы нечёткого поиска:

- семантические сети;
- регулярные выражения;
- нечёткий поиск на основе метрик;
- нейронные сети;
- хеширование по сигнатуре;
- метод N-грамм;
- BK-деревья;

- фонетические алгоритмы.

Методы точного поиска:

- поиск по словарю;
- полнотекстовый поиск (суффиксные деревья).

Многие из этих методов хорошо разработаны и имеют широкое применение в различных прикладных задачах [94, 146, 98, 158].

В маркетинговых коммуникациях одним из наиболее важных типов данных о клиенте – это его контактные данные, например, телефон или e-mail. Извлечение из текста такой записи, как email не является сложной задачей. Это можно выполнить с использованием регулярных выражений. Извлечение телефонов может быть сопряжено с рядом трудностей, обусловленных наличием в тексте цифр, которые могут относиться к другим сущностям, например, датам, номеру дома и т.д. Например, последовательность «(456)7-51-52» с большей вероятностью будет представлять собой номер телефона, а «05.13-2011» – дату.

Распознавание необходимых сущностей с достаточной точностью не всегда достижимо с помощью перечисленных методов [94, 98], для них зачастую характерна низкая производительность (большие затраты времени или памяти) [146, С. 43], что сильно затрудняет обработку объёмов данных. Кроме того, эти методы разработаны для нечёткого сравнения одних и тех же сущностей, например, когда производится поиск опечаток. Они позволяют ответить на такой вопрос, как «имеется ли в анализируемой последовательности опечатка?» или «насколько похожи две анализируемые последовательности?», но в большинстве случаев эти методы не дают ответа на вопрос типа «чем является данная анализируемая последовательность? Фамилией? Телефоном? E-mail-ом? Или бесполезным набором символов?»

Для ответа на подобные вопросы в данном исследовании предлагается алгоритм, который позволяет произвести нечёткий поиск представителей интересующего класса сущностей по некоторому набору эталонных сущностей. Для этого предлагается перестроить метод нечёткого сравнения

так, чтобы он находил совпадения по характерным признакам классов. Наиболее удобным методом представляется метод вычисления метрики Дамерау-Левенштейна.

Метрика Левенштейна используется для того, чтобы измерить расстояние между двумя символьными последовательностями. Метрика Левенштейна – это число, характеризующее минимальное количество операций редактирования: удаления, замены или вставки символов для превращения одной строковой последовательности в другую. Эта метрика предложена Владимиром Левенштейном в 1965 году и названа в честь её автора [32].

Метрику Левенштейна иначе можно называть расстоянием редактирования. Она относится к семейству метрик строковых расстояний [128, С. 21] и имеет много общего с алгоритмами выравнивания последовательностей.

Формально метрику Левенштейна для строк s1, s2, $L_{s1,s2}$ можно задать следующим образом (29):

$$L_{s1,s2}(i,j) = \begin{cases} \max(i,j), \min(i,j) = 0, \\ min \begin{cases} L_{s1,s2}(i-1,j) + 1 \\ L_{s1,s2}(i,j-1) + 1 \\ L_{s1,s2}(i-1,j-1) + 1, \text{если } a_i \neq b_j \end{cases} \end{cases} \quad (29)$$

Следует заметить, что первый элемент при вычислении минимума относится к удалению, второй относится к вставке, а третий отражает совпадение или несовпадение последовательностей.

Дать общую характеристику для метрики Левенштейна возможно на основе следующих умозаключений:

1. Нижнюю оценку расстояния Левенштейна можно найти, вычислив разницу в длине символьных последовательностей. Такую операцию можно проделать для любых строк.

2. Верхнюю оценку расстояния Левенштейна можно вычислить, найдя длину наибольшей последовательности.

3. Если последовательности идентичны, расстояние равно нулю.

4. Верхнюю оценку расстояния Левенштейна также даёт Метрика Хемминга [101].

5. По свойству метрик «неравенство треугольника» расстояние Левенштейна между двумя символьными последовательностями не будет превышать сумму расстояний этих последовательностей до третьей.

Метрика Хемминга отражает количество различающихся символов в соответствующих позициях строковых последовательностей. Если имеются последовательности «БГДЗ» и «АБГД», то расстояние Хемминга для них будет равно четырём, поскольку разные символы находятся на первой, второй, третьей и четвёртой позиции соответственно. Расстояние Левенштейна для этих же последовательностей будет равно двум, так как необходимо выполнить две операции: удалить букву «А» и вставить букву «З».

В текстах, где ожидается незначительное число различий можно применять Метрику Левенштейна для нечёткого сравнения строк. В некоторых случаях короткие последовательности можно брать из словаря. Метрика Левенштейна имеет широкую сферу применений: системы проверки правописания, системы коррекции оптически распознанного текста, системы поддержки перевода с естественных языков на базе отрывков текста, переведённых ранее.

Сложность вычисления метрики Левенштейна оценивается, как O(m*n), где m, n – длины строковых последовательностей.

Перечислим существующие алгоритмы вычисления метрики Левенштейна.

Рекурсивный алгоритм. Этот алгоритм непосредственно отражает логику, заложенную в его формальном описании. Он удобен для понимания, однако является неэффективным, так как метрика Левенштейна вычисляется по нескольку раз для одних и тех же строк.

Алгоритм Вагнера-Фишера [155]. В данном алгоритме используется матрица расстояний между всеми префиксами анализируемых последовательностей. На основе уже вычисленного расстояния вычисляется расстояние каждого следующего префикса до предыдущего, рисунок 15.

		А	В	О	О	Г	И
	0	1	2	3	4	5	6
Е	1	1	2	3	4	5	6
В	2	2	1	2	3	4	5
О	3	3	2	1	2	3	4
О	4	4	3	2	1	2	3
В	5	5	4	3	2	2	3
И	6	6	5	4	3	3	2
Ж	7	7	6	5	4	4	3

		К	У	П	А	Л	Д	О	Н
	0	1	2	3	4	5	6	7	8
К	1	0	1	2	3	4	5	6	7
А	2	1	1	2	2	3	4	5	6
Р	3	2	2	2	3	3	4	5	6
Д	4	3	3	3	3	4	3	4	5
О	5	4	3	4	4	4	4	3	4
Н	6	5	4	4	5	5	5	4	3

Источник: составлено автором.

Рисунок 15 – Результат работы алгоритма Вагнера-Фишера на двух примерах

Расстояние Левенштейна посчитано в правом нижнем элементе матрицы.

Итеративный алгоритм с двумя столбцами матрицы дан на рисунке 16.

```
int CalculateLevenshteinMetric(string s1, string s2)
{
    if (s1 == s2) return 0;
    if (s1.Length == 0) return s2.Length;
    if (s2.Length == 0) return s1.Length;

    int[] D1 = new int[s2.Length + 1];
    int[] D2 = new int[s2.Length + 1];

    for (int i = 0; i < D1.Length; i++)
        D1[i] = i;
    for (int i = 0; i < s1.Length; i++)
    {
        D2[0] = i + 1;
        for (int j = 0; j < s2.Length; j++)
        {
            int c = (s1[i] == s2[j]) ? 0 : 1;
            D2[j + 1] = Math.Min(Math.Min(D2[j] + 1, D1[j + 1] + 1), D1[j] + c);
        }
        for (int j = 0; j < D1.Length; j++)
            D1[j] = D2[j];
    }
    return D2[s2.Length];
}
```

Источник: составлено автором.

Рисунок 16 – Реализация итеративного алгоритма с двумя столбцами матрицы

Этот алгоритм основывается на алгоритме Вагнера-Фишера, но он более эффективен, так как для вычисления используются всего два столбца матрицы. В общем случае итеративный алгоритм с двумя столбцами матрицы эффективнее предыдущего по требуемой памяти.

2.1.1 Метрика Дамерау-Левенштейна

Метрика Дамерау-Левенштейна представляет собой расширение метрики Левенштейна.

Метрика Дамерау-Левенштейна – это расстояние между двумя конечными символьными последовательностями, которое соответствует наименьшему количеству операций удаления, замены, вставки и транспозиции соседних символов, необходимому для превращения одной символьной последовательности в другую. [69, 60]

Более 80% ошибок набора текста человеком устраняется перечисленными операциями редактирования. Дамерау в своей работе рассматривал те ошибки, которые возможно исправить с помощью только одной операции редактирования [78].

Формально метрику Дамерау-Левенштейна для строк s1, s2, $DL_{s1,s2}$ можно задать следующим образом (30):

$$DL_{s1,s2}(i,j) = \begin{cases} \max(i,j), \min(i,j) = 0, \\ \min \begin{cases} DL_{s1,s2}(i-1,j)+1 \\ DL_{s1,s2}(i,j-1)+1 \\ DL_{s1,s2}(i-1,j-1)+1, a_i \neq b_j \\ DL_{s1,s2}(i-2,j-2)+1 \end{cases} i,j > 1, a_i = b_{j-1} \\ \min \begin{cases} DL_{s1,s2}(i-1,j)+1 \\ DL_{s1,s2}(i,j-1)+1 \\ DL_{s1,s2}(i-1,j-1)+1, a_i \neq b_j \end{cases} \end{cases}, \quad (30)$$

где

$DL_{s1,s2}(i-1,j)+1$ – удаление

$DL_{s1,s2}(i,j-1)+1$ – вставка

$DL_{s1,s2}(i-1,j-1)+1, a_i \neq b_j$ оператор проверки последовательности на совпадение

$DL_{s1,s2}(i-2,j-2)+1$ транспозиция

Введение операции транспозиции сопряжено с дополнительной сложностью.

Для стоимости операции транспозиции существует следующее ограничение (31) [120, С. 180]:

$$2W_T \geq W_I + W_D \tag{31}$$

Имеется возможность вычислить Метрику Дамерау-Левенштейна также с помощью модифицированного алгоритма Вагнера-Фишера. В этом случае оценка сложности составит O(m*n), где m, n – длины символьных последовательностей.

В частности, Метрика Дамерау-Левенштейна широко используется в обработке текстов, а также в вычислении последовательностей молекул дезоксирибонуклеиновой кислоты.

Метрику Дамерау-Левенштейна можно преобразовать для поиска классов сущностей.

В прикладных задачах может потребоваться не столько ответ на вопрос «является ли символьная последовательность А похожей на символьную последовательность В», сколько ответ на вопрос «относится ли символьная последовательность А к тому же классу, что и символьная последовательность В». Например, если известно, символьная последовательность может содержать важные данные, такие, как номер дома, телефон, дата, необходимо выяснить, как можно распознать эти данные.

Имеется символьная последовательность: «Советская, д.5, Поляков Василий Викторович, тел.: +7(985)345-55-32, родился 26.08.1974». Необходимо разработать такой алгоритм, который с наиболее высокой точностью распознавал вложенные последовательности и относил бы их к нужному классу. Для этого предлагается воспользоваться метрикой Дамерау-

Левенштейна с матрицей стоимостей замены символов. Пример такой матрицы представлен в таблице 6. Эту матрицу необходимо ввести для того, чтобы лучше отделить символьные последовательности, для которых характерны отдельные символы от остальных символьных последовательностей. Аналогично можно задать веса вставки и удаления очень важных или, наоборот, малозначимых символов [39, С. 104].

Таблица 6 – Матрица стоимостей замены символов. \d – цифра, \w – буква

pattern	\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.

Введение матрицы стоимостей замены символов несколько увеличивает сложность алгоритма за счёт поиска значений в матрице (32):

$$O'(m * n * \log k) \quad (32)$$

где m и n – длины символьных последовательностей, k – размер матрицы стоимостей, $2\log_2 k$ – наибольшее количество операций для поиска значения стоимости замены пары символов k1, k2.

Далее необходимо составить, по крайней мере, не меньше одной эталонной последовательности – представителя класса.

Представители класса телефонов:

[+7(985)-665-43-55];

[8(985)6654355];

[9856654355].

Представители класса дат:

[2015-01-07];

[9.01.1993]

Составим таблицу, где укажем расстояния Дамерау-Левенштейна с учётом весов, где названиям строк соответствуют подпоследовательности в анализируемой строке, а названиям столбцов соответствуют представители классов в таблице 7.

Таблица 7 - Матрица расстояний Дамерау-Левенштейна с произвольными весами операций подстановки

Подпоследо-вательность	**Представители классов**				
	9.01.1993	2015-01-07	98566543	8(985)6654355	+7(985)665-43-55
д.5	8	10	10	13	17
Советская	9	10	10	13	17
26.08.1974	1	4	4	6	10
Родился	9	10	10	13	17
+7(985)345-55-32	11	6	6	3	0
тел.:	8	10	10	13	17
Викторович	10	10	10	13	17
Василий	9	10	10	13	17
Поляков	9	10	10	13	17

Источник: составлено автором.

После чего нужно выбрать порог отсечения для значения метрики, по которому представители классов будут считаться различными. В анализируемом случае можно заметить, что максимальное из расстояний между разными представителями класса равно 6. Одним из вариантов выбора порога отсечения может быть вариант с использованием принципа минимальных потерь. При использовании такого способа порог отсечения в данном случае равен шести.

Необходимо для каждой последовательности найти такого представителя класса, до которого расстояние Дамерау-Левенштейна будет наименьшим и при этом не будет превышать порог отсечения.

В исследуемом примере вложенные последовательности были верно отнесены к своим классам. В отдельных случаях требуется тонкая настройка

весов матрицы стоимостей. Немаловажную роль играет правильный выбор эталонных представителей классов.

Из базы данных лизинговой компании обработано 71638 записей, полученных из открытых источников. Результаты извлечения данных представлены в таблице 8.

Таблица 8 – Результаты анализа записей

Наименование параметра	**Значение**
Всего записей	71638
Записей, содержащих цифры	69834
Количество извлечённых корректных телефонов	54769
Количество записей, содержащих символ «@»	24798
Количество извлечённых корректных e-mail	23998
Доля ложно распознанных телефонов без применения матрицы стоимостей (выборочная проверка случайных 100 записей)	0.14
Доля ложно распознанных телефонов с применением матрицы стоимостей (выборочная проверка случайных 100 записей)	0.3
Дозваниваемость полученных телефонов	0.78
Время работы, сек	6428

Источник: составлено автором.

Практический пример показывает, что использование матрицы стоимостей для нечёткого сравнения с вычислением метрики Дамерау-Левенштейна может оказать положительное влияние на правильность распознавания сущностей искомых классов для некоторых прикладных задач. Приведённый подход имеет ограничения. В частности, размер матрицы стоимостей негативно влияет на производительность алгоритма. Помимо этого, настройка матрицы весов сильно зависит от набора данных. Это ограничивает универсальность предложенного подхода. Вместе с тем при пакетной обработке данных разработанный инструмент является гибким и удобным в настройке.

В отличие регулярных выражений метрика Дамерау-Левенштейна обладает таким свойством, что не требуется рассматривать все возможные

комбинации данных заданного типа. В случае выбора метрики Дамерау-Левенштейна можно задать допустимое отклонение от заданного шаблона.

В виду отсутствия достаточной и репрезентативной информации о том, какие записи на самом деле являются дублирующимися, объективно оценить качество дедупликации невозможно. Для тестирования можно воспользоваться косвенными признаками, например, наличием общих контактных данных у юридических лиц, признанных дублирующимися. Будем считать, что у юридических лиц, верно признанных дублями, вероятность наличия общих контактных данных выше, чем у дублей, признанных таковыми неверно. Тогда введём критерий качества дедупликации, который представляет собой отношение вероятностей нахождения общих контактных данных клиентов из одной группы и их нахождения у клиентов из разных групп (33, 34, 35, 36) [38]:

$$DQ = \frac{\sum_{i=1}^{n} P_{same}(x_i)}{1 + \sum_{i=1}^{n} P_{diff}(x_i) * n}, \tag{33}$$

$$P_{same}(x) = \frac{\sum_{j=1}^{l} sign(x = x_j)}{l - 1}, x \in G_k, x_j \in G_k \backslash x, \tag{34}$$

$$P_{diff}(x) = \frac{\sum_{j=1}^{n-l} sign(x = x_j)}{n - l}, x \in G_k, x_j \notin G_k, \tag{35}$$

$$sign(x = x_j) = \begin{cases} 1, x = x_j \\ 0, x \neq x_j \end{cases}, \tag{36}$$

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.

Критерием качества извлечения данных может служить отношение количества данных, извлечённых верно, к количеству данных, извлечённых неверно (37):

$$EQ = \frac{TP}{1 + FP} \tag{37}$$

Записи о клиентах лизинговой компании были дедуплицированы различными методами. Статистика дедупликации с помощью адаптированной метрики Дамерау-Левенштейна приведена в таблице 9.

Таблица 9 – Статистика дедупликации

Количество в группе	**Количество групп**
2	82517
3	11583
4	4836
5-6	4495
7-10	2986
11-15	1368
16-20	615
21-30	525
31-50	375
51-70	131
71-100	89
101-150	64
151-200	26
200+	23

Источник: составлено автором.

Статистика извлечения и проверки телефонных номеров приведена в таблице 10.

Таблица 10 – 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 11 приведены показатели качества дедупликации и извлечения данных с применением различных методов.

Таблица 11 – 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.

2.2 Разработка составного ключа базы данных для оптимизации индексированного поиска

При обработке больших объёмов данных возникают проблемы, связанные с их упорядочиванием и поиском. Индексы – основной инструмент, позволяющий быстро находить записи по ключу. Однако, в некоторых случаях задача построения индекса носит нетривиальный характер (например, когда имеются пропуски в данных) [35, С. 105]. Пример решения одной из таких задач рассмотрен в данном разделе.

2.2.1 Назначение индексов

Часто для поиска информации о сущностях из одной таблицы требуется обращаться к другим таблицам. Таблицы могут содержать большой объём записей, при этом время поиска увеличивается мультипликативно. Для того, чтобы найти всю информацию о сущностях из таблицы A в таблице B, нужно для каждой сущности из таблицы A перебрать все записи в таблице B. Время поиска в таком случае будет равно (38):

$$T = t * N_A * N_B, \quad (38)$$

где t – время, требуемое на выбор одной записи, N_A – количество записей в таблице A, N_B – количество записей в таблице B.

В общем случае время цепного поиска (где в каждой следующей таблице содержится информация о сущностях предыдущей) составит (39):

$$T = t * \prod N_i \quad (39)$$

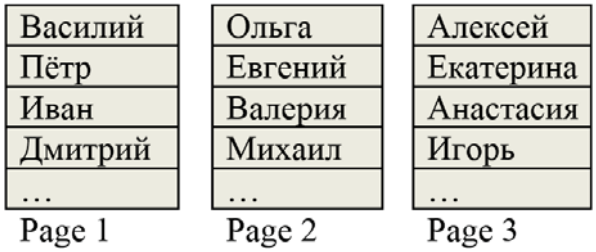
Таким образом, при одинаковом количестве строк в таблицах время поиска растёт экспоненциально (N^i). Как правило, с каждой последующей таблицей в цепи количество записей растёт.

Индекс – это объект в базе данных, предназначенный для оптимизации поиска данных. [29]

Поиск по индексированным ключам может оказаться более эффективным. Простейшим примером индекса может служить оглавление книги. Вместо перебора всех параграфов, читатель может открыть нужный параграф, зная номер его страницы.

При создании таблицы данные хранятся неупорядоченно (в куче) на страницах по 8 Кб, рисунок 17.

Если необходимо выбрать записи, удовлетворяющие условию, показанному на рисунке 18, SQL-сервер прочитает все записи в таблице и выведет те, которые удовлетворяют заданному условию. У сервера нет информации о том, что в таблице есть только одна такая запись, пока не заданы ограничения уникальности или первичного ключа.



Источник: составлено автором.

Рисунок 17 – Пример страниц

```
SELECT *
FROM Persons
WHERE Name = ‘Михаил’
```

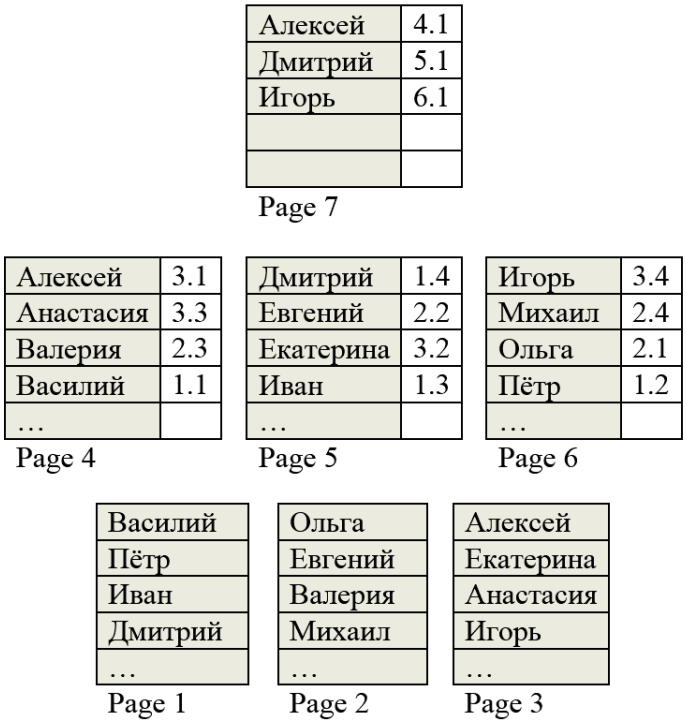
Источник: составлено автором.

Рисунок 18 – Запрос на выборку по полю «Name»

Индексы хранятся в виде сбалансированных деревьев. Это значит, что длина пути ветвления постоянна. Простроим индекс для таблицы Persons, рисунок 19.

Указатель p.n хранит адрес записи с номером n на странице p. Теперь для нахождения записи ‘Михаил’ необходимо перебрать шесть записей, вместо двенадцати: {7.1; 7.2; 7.3; 6.1; 6.2; 2.4}.

Индексы можно создавать как на отдельные поля, так и на набор полей. Если необходимо быстро выбрать записи, удовлетворяющие условию на рисунке 20, нужно создать индекс совместно на поля “Brand” и “Price”, рисунок 21.



Источник: составлено автором.

Рисунок 19 – Пример индекса

```
SELECT *
FROM Automobile
WHERE Brand = 'Skoda' and Price = 1000000
```

Источник: составлено автором.

Рисунок 20 – Запрос на выборку по полям «Brand» и «Price»

```
CREATE NONCLUSTERED INDEX INC_A on Automobile( Brand ASC, Price ASC )
```

Источник: составлено автором.

Рисунок 21 – Запрос на создание кластерного индекса

Порядок сортировки обоих полей указан по возрастанию, хотя, в данном случае это не важно. Выше был приведён пример создания некластерного индекса. Таких индексов можно создать неограниченное число. Кластерный индекс аналогичен некластерному с тем отличием, что на уровне его листьев находятся не упорядоченные ссылки на записи в таблице, а упорядоченные записи таблицы. Так как упорядочить строки в таблице можно одновременно только одним способом, кластерный индекс может быть только один. Очевидно, что преимуществом кластерного индекса является более высокая скорость доступа к данным.

Таким образом, время поиска записи в таблице A может сократиться с N_Aдо $\log_2 N_A$ (время бинарного поиска), а время цепного поиска составит (40):

$$T = t * \prod \log_2 N_i \tag{40}$$

При одинаковом количестве строк в таблицах время поиска составит (41):

$$T = t * n * \log_2 N, \tag{41}$$

где n – количество таблиц в цепи.

2.2.2 Формирование ключа

В случае, когда необходимо искать значения в текстовых полях, задача осложняется тем, что нужно проверить достаточно большое число байт, которое зависит от объёма текста. Это может замедлить поиск.

Для быстрого поиска по тексту можно использовать хеширование [37, С. 18].

Хеширование – отображение последовательности бит переменной длины в последовательность бит фиксированной длины с помощью хеш-фукнций (функций свёртки). [93, 13]

Для составления индекса возьмём 128-битный алгоритм хеширования MD5. Каждому текстовому полю поставим в соответствие свой хеш – «достаточно уникальный» ключ, таблица 12. Хеш-код MD5 занимает 128 бит в памяти, в то время, как слово из 10 символов в UNICODE занимает 160 бит. Каждый новый символ добавляет ещё 16 бит.

Таблица 12 – Таблица хешей

Наименование	**Хеш**
Василий	86601e5039ba231c5e529f29c56f86c2
Пётр	ff61f06f73e1868493ec04a258f8149b
Иван	acd41b5fc27242d19c244185ba6732f2
Дмитрий	33b67031842d4c56d9dc03571081b77d
…	…

Источник: составлено автором.

MD5 в нашем случае будет ключом фиксированной длины при этом, можно сохранить значительное количество памяти, составив словарь соответствия строковых значений своему ключу.

Большую роль хеш-сумма может сыграть в условиях неполной информации.

Например, если стоит задача найти совпадения по фамилии, имени и отчеству в двух таблицах, причём, если в одной из таблиц какое-либо из

полей неизвестно, то сравнение с аналогичным полем другой таблицы вернёт истину.

Например, записи с наличествующим и отсутствующим отчеством в таблице 13 будем считать эквивалентными. Тогда, чтобы ускорить поиск, присвоим каждому полю свой ключ, таблица 14.

Таблица 13 – Пример «эквивалентных» записей по ФИО

LastName	FirstName	PatronymicName
Пеклин	Александр	Михайлович

LastName	FirstName	PatronymicName
Пеклин	Александр	NULL

Источник: составлено автором.

Таблица 14 – Таблица хешей фамилий, имён и отчеств

LastName	LNMD5	FirstName	FNMD5	Patronymic Name	PNMD5
Пеклин	76a033e832dd0be55aaf6766a681afae	Александр	3d801da09e7d82668e226799d9db91dc	Михайлович	6dd133fe6a47c6f3249ea86590639811

LastName	LNMD5	FirstName	FNMD5	Patronymic Name	PNMD5
Пеклин	76a033e832dd0be55aaf6766a681afae	Александр	3d801da09e7d82668e226799d9db91dc	NULL	NULL

Источник: составлено автором.

Для индексации незаполненных полей (со значениями «NULL») можно воспользоваться следующим алгоритмом:

1. каждое поле MD5 разделить на два поля (MD5_begin, MD5_end) и поместить в оба поля значения MD5;

2. если MD5 принимает значение NULL, в поле MD5_begin поставить минимальное значение 00000000000000000000000000000000, а в поле MD5_end поставить максимальное значение ffffffffffffffffffffffffffffffff;

3. проиндексировать таблицу по всем полям (MD5_begin, MD5_end).

Таблица 15 – Соединение по ключу

LMD5	LMD5_begin	LMD5_end	…	PMD5	PMD5_begin	PMD5_end
bbc45d47 dc540df0 6b7f40ba 081e0fb1	bbc45d47dc5 40df06b7f40 ba081e0fb1	bbc45d47dc 540df06b7f 40ba081e0f b1	…	NULL	00000000000 00000000000 0000000000	fffffffffffff fffffffffffff ffffff

Источник: составлено автором.

Теперь искать записи можно следующим образом:

Совпадение считается выполненным для строк из таблиц A и B, если одновременно выполняются условия:

A.LMD5_begin <= B.LMD5_end

A.LMD5_end >= B.LMD5_begin

A.FMD5_begin <= B.LMD5_end

A.FMD5_end >= B.LMD5_begin

A.PMD5_begin <= B.LMD5_end

A.PMD5_end >= B.LMD5_begin

Аналогичные преобразования можно провести для поиска изменившихся связей между сущностями различных баз данных. Например, если в двух базах данных 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 16.

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 запись из таблицы 17, так и новую запись из таблицы 18.

Таблица 16 – 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.

Таблица 17 – Пример найденной старой записи

A_id	**B_id**
NULL	6111

Источник: составлено автором.

Таблица 18 – 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.

ГЛАВА 3
РАЗРАБОТКА МАТЕМАТИЧЕСКОЙ МОДЕЛИ ОТКЛИКА НА МАРКЕТИНГОВЫЕ КОММУНИКАЦИИ

3.1 Построение отображения клиентских характеристик на нелинейное двумерное многообразие. Кластеризация клиентской базы

Для повышения эффективности поиска закономерностей, отбора целевых признаков, формального обоснования целесообразности их внедрения в регрессионную модель, а также интерпретации признаков для поддержки принятия решений, используются различные методы визуализации данных. В данном разделе приводится исследование отклика на маркетинговые коммуникации с применением различных подходов: простые статистические методы, упругие карты, а также самоорганизующихся карты Кохонена. В ходе проведённого анализа были выявлены основные ключевые характеристики и закономерности, непосредственно оказывающие влияние на конверсию в маркетинговых коммуникациях. Результаты исследования могут быть использованы для принятия управленческих решений, а также для построения регрессионной модели отклика на маркетинговые коммуникации.

3.1.1 Простейшие методы анализа маркетинговых коммуникаций

Конверсия маркетинговых коммуникаций является целевым оптимизируемым показателем. Этот показатель вычисляется как отношение числа участников коммуникации, совершивших целевое действие к их общему числу (42) [30]:

$$Conv_i = \frac{TS_i}{TS_i + TF_i}, \tag{42}$$

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

Используя систему простых показателей можно графически представить значимость признаков по отношению к конверсии.

Доля клиентов, обладающих выбранным признаком выражается как (43):

$$P_i = \frac{V_i}{V}, \tag{43}$$

где V_i – количество клиентов, имеющих признак i, V – общее число клиентов в группе, P_i – доля клиентов, имеющих признак i.

Распространённость признака в целевой группе подразумевает долю целевых клиентов, обладающих данным признаком (44):

$$S_i = \frac{p_i}{TS}, \tag{44}$$

где p_i – число клиентов в целевой группе, имеющих признак i, TS – число клиентов в целевой группе (иначе говоря, удачных попыток коммуникации), S_i – доля клиентов, имеющих признак i в целевой группе.

Можно, таким образом, вывести коэффициент характерности признака, отражающий сравнительную степень выраженности этого признака для целевых клиентов по сравнению со всей группой (45):

$$\mathrm{C}_i = \frac{S_i - P_i}{P_i} = \frac{\frac{p_i}{TS} - \frac{V_i}{V}}{\frac{V_i}{V}} = \frac{V \times p_i}{V_i \times TS} - 1, \tag{45}$$

где C_i – коэффициент характерности признака i у целевой группы. Этот коэффициент численно выражает важность обладания информацией о факте наличия признака i.

Ниже описан портрет потенциального лизингополучателя, который был составлен по результатам проведения маркетинговой кампании, с помощью полученных выше метрик, рисунок 22.

На рисунке 23 показаны синтетические признаки, состоящие из совокупности условий, которые были выявлены на ряду с основными.

Признак	Условие	S	P	C
Количество проектов (интерес)	[0 - 17]	88%	97%	-0,09
Дней с последней покупки	[0 - 190]	86%	59%	0,46
Дней между покупками группа 1	[476 - inf)	59%	15%	2,90
Дней между покупками группа 2	[19 - 77]	12%	2%	4,83
Дней между покупками группы 1 и 2	[19 - 77] U [476 - inf)	71%	17%	3,14
Времени с последнего проекта (интереса)	[0 - 128]	82%	64%	0,28
Дней между проектами	[1075 - 2182]	49%	26%	0,85
Покупал более 1 раза	ИСТИНА	100%	51%	0,97
Пришёл от дилера	ИСТИНА	61%	52%	0,19
Пришёл с сайта	ИСТИНА	90%	92%	-0,02
Категория по выручке группа 1	[8,5М - 34М)	29%	14%	1,05
Категория по выручке группа 2	[0 - 0,85М)	22%	40%	-0,44
Категория по выручке группы 1 и 2	[0 - 0,85М) U [8,5М - 34М)	51%	54%	-0,06
Находится в Москве	ЛОЖЬ	100%	90%	0,11
Средний возраст транспортных средств	[0 - 11]	96%	89%	0,08
Работает в разных отраслях	ИСТИНА	96%	96%	0,00

Источник: составлено автором.

Рисунок 22 – Анализ основных характеристик клиента

Синтетический признак	Условие	S	P	C
Дней с последней покупки	[0-146]	45%	30%	0,4966
Пришёл от дилера	ИСТИНА			
Средний возраст транспортных средств	[0-11]			
Дней между покупками		71%	36%	0,97921862
Совершено более двух покупок	ИСТИНА			

Источник: составлено автором.

Рисунок 23 – Анализ синтетических характеристик клиента

Простой анализ показывает, что самая ценная информация о клиенте (для определения его отклика на маркетинговые коммуникации) выражается следующим образом «в среднем клиент совершает покупки раз в 19-77 дней», но таких клиентов всего 2%.

Данные о выручке клиента, которая составляет менее 850 000 рублей, свидетельствуют о том, что, клиент, скорее всего, не склонен к покупке. Данный признак заполнен у 40% клиентов, что можно считать относительно неплохим показателем.

Точный портрет целевого клиента будет записываться следующим образом «клиент с годовой выручкой 8,5 – 34 млн. руб., часто совершающий покупки, обратился через дилера, недавно интересовался или заключал

сделки с компанией». Как правило, подобных клиентов или очень мало, или не существует вовсе. Тем не менее, такой портрет позволяет дать ориентиры для стратегического планирования и подготовки очередных маркетинговых кампаний.

3.1.2 Применение нелинейных методов отображения данных в анализе маркетинговых коммуникаций

В настоящее время существуют и активно используются методы, позволяющие визуально представить структуру многомерного набора данных целиком.

Визуализация данных представляет собой метод отражения многомерного массива данных на двумерной плоскости, с сохранением, хотя бы качественно, основных пропорций исходного распределения. Ими могут являться такие виды отношений, как топология, кластерная структура, различные зависимости между признаками, позиции точек данных в многомерном пространстве и т.д. [25, С. 48].

Применение нелинейных методов визуализации используется с целями:

- описания закономерностей в данных;
- восстановления пропущенных значений;
- сжатия информации;
- наглядного геометрического представления данных;
- прогнозирования и построения регрессионных моделей.

Традиционно решение поставленных задач решается методами целенаправленного проецирования и многомерного шкалирования. Поиск отображения данных многомерного пространства на двумерную плоскость путём оптимизации функционала от координат точек данных производится в методах целенаправленного проецирования [2]. При решении задач методами многомерного шкалирования имеется информация только о расстоянии между точками данных, но отсутствует исходная информация об их

координатах. Задача шкалирования сводится к поиску координат точек, удовлетворяющих матрице расстояний [3, 19, 47].

Сравнительно новый метод визуализации данных представляют собой самоорганизующиеся карты Кохонена [112], также отечественными исследователями в Институте Вычислительного Моделирования г. Красноярска разработан метод упругих карт [26]. Математический аппарат этих методов ориентирован на поиск оптимальной ориентации вложенных поверхностей в структуре многомерных данных.

Метод самоорганизующихся карт Кохонена описывает проецирование точек данных на сетку узлов, пошагово приближенных к скоплениям точек, при минимизации ошибки аппроксимации (46):

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - y_{BMU}(X_i)\right)^2}, \quad (46)$$

где $y_{BMU}(X_i)$ – ближайший узел к точке X_i.

Узлы сетки генерируются и сдвигаются в направлении к точкам данных по следующему алгоритму (47):

$$(y_j)' = y_j + h(r(y_j, y_{BMU}), t)(X_i - y_i), j = 1 \dots p, \quad (47)$$

где $h(x,t)$ – функция соседства узлов, $r(y1, y2)$ – расстояние между узлами сетки $y1$ и $y2$, p – число узлов сетки.

Наиболее часто используемые функции соседства:

Гауссова функция (48):

$$h(x,t) = \alpha(t)e^{-\frac{x^2}{2\sigma^2(t)}} \quad (48)$$

Bubble-функция (49):

$$h(x,t) = \begin{cases} \alpha(t), x \le \sigma(t) \\ 0, x > \sigma(t) \end{cases}, \quad (49)$$

где $\sigma(t)$ – радиус захвата соседей, $\alpha(t)$ – т.н. темп обучения [25].

Выполним нормализацию распределения данных по формуле (50):

$$X_i \to X_i' | x_{ij}' = x_{F(x_{ij})}, j \in n, \quad (50)$$

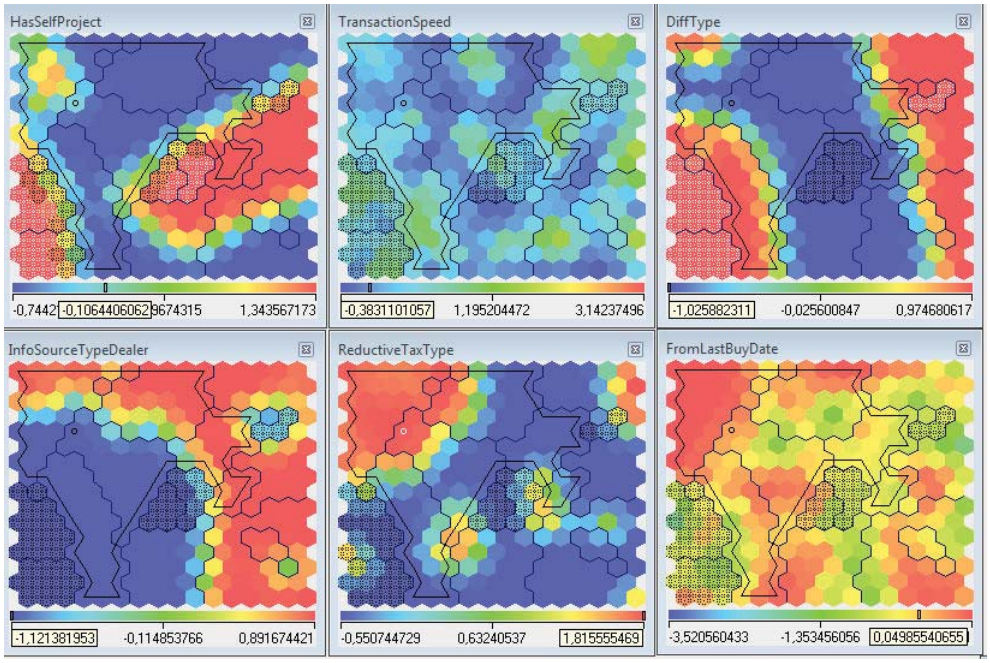
где n – объём данных, $x_{i1}, \dots, x_{in}$ – уникальные значения признака $X_i, x_{i1}', \dots, x_{in}'$ – уникальные значения признака X_i', $F(x_{ij})$ – эмпирическая функция распределения признака X_i (51, 52):

$$F(x_{ij}) = \overline{P(x < x_{ıj})} = P(x < x_{ij}) + \frac{P(x = x_{ij})}{2}, \quad (51)$$

$x_{F(x_{ij})}$ – квантиль уровня $F(x_{ij})$ для нормального распределения.

$$F(x) = N = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \quad (52)$$

При построении карты Кохонена будем по очереди отбрасывать неинформативные признаки, рисунок 24. Закрашенная область (2, 3, 4) выделяет высококонверсионные кластеры, область, обведённая контуром (0, 1, 6, 7, 9) – низкоконверсионные.



Источник: составлено автором.

Рисунок 24 – Карта Кохонена в разрезе исследуемых признаков

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) [34, С. 204].

С помощью обратного преобразования признаков можно вывести их настоящие диапазоны, соответствующие нормированным значениям.

Номер кластера	Y=1	Количество	Конверсия
3	48	408	11,8%
2	87	812	10,7%
4	16	180	8,9%
5	34	803	4,2%
10	78	1912	4,1%
8	61	1513	4,0%
6	21	704	3,0%
9	30	1075	2,8%
0	21	870	2,4%
7	31	1363	2,3%
1	24	1183	2,0%

Источник: составлено автором.

Рисунок 25 – Статистика по кластерам

Группа кластеров	%Количество	%(Y=1)	Средняя конверсия	Превышение средней конверсии
С высокой конверсией	13%	33,5%	10,5%	105%
С низкой конверсией	48%	28,2%	2,5%	-51%

Источник: составлено автором.

Рисунок 26 – Статистика по группам кластеров

Метод упругих карт представляет собой обобщение метода главных компонент. Построение вложенного многообразия является оптимизационной задачей. Она заключается в нахождении нелинейной (упругой) поверхности с минимальным искажением дисперсии проекций точек данных по сравнению с дисперсией начальных точек данных. Упругую

сетку изначально можно расположить в плоскости первых двух главных компонент.

Формулировка задачи для метода главных компонент звучит следующим образом:

Начальные признаки должны линейно восстанавливаться по новым с наименьшей погрешностью (53, 54):

$$\hat{f}_j(x) = \sum_{s=1}^{m} g_s(x) u_{js}, j = 1, \dots, n, \forall x \in X \tag{53}$$

как можно точнее на обучающей выборке $x_1, \dots, x_l$:

$$\sum_{i=1}^{l} \sum_{j=1}^{n} (\hat{f}_j(x_i) - f_j(x_i))^2 \to \min_{\{g_s(x_i)\},\{u_{js}\}}, \tag{54}$$

где $g_1(x), \dots, g_m(x)$ – новые числовые признаки, $m \leq n$, $f_1(x), \dots, f_n(x)$ – начальные числовые признаки.

Начальная матрица «объекты-признаки» (55):

$$F = \begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_l) & \cdots & f_n(x_l) \end{pmatrix} \tag{55}$$

Преобразованная матрица «объекты-признаки» (56):

$$G = \begin{pmatrix} g_1(x_1) & \cdots & g_m(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_l) & \cdots & g_m(x_l) \end{pmatrix} \tag{56}$$

Матрица линейного отображения признаков (57):

$$U = \begin{pmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nm} \end{pmatrix} \tag{57}$$

Матричная запись линейного отображения (58):

$$\hat{F} = GU^T \approx F \tag{58}$$

Матричная запись оптимизируемого функционала (59):

$$\sum_{i=1}^{l} \sum_{j=1}^{n} (\hat{f}_j(x_i) - f_j(x_i))^2 = \|GU^T - F\| \to \min_{G,U} \tag{59}$$

Решить задачу метода главных компонент – значит найти матрицу U. Столбцы U являются собственными векторами матрицы $F^T F$, соответствующими наибольшим собственным значениям $\lambda_1, \dots, \lambda_m$ [16] (60).

$$G = FU \tag{60}$$

Узлы упругой сетки пронумерованы индексами $y^{ij}, i = 1 \dots p, j = 1 \dots q$. Каждому узлу y^{ij} соответствует подмножество точек данных $K_{ij} (i = 1 \dots p, j = 1 \dots q)$ при этом для каждой точки из этого подмножества этот узел ближайший (61):

$$K_{ij} = \{x \in X | \left\| y^{ij} - x \right\|^2 \to \min_{i,j} \quad \} \tag{61}$$

Конечная конфигурация упругой сетки настраивается путём оптимизации линейной комбинации функционалов, обозначающих следующие характеристики:

1) приближение к точкам данных (62):

$$D_1 = \sum_{ij} \sum_{X_k \in K_{ij}} \left\| X_k - y^{ij} \right\|^2 \tag{62}$$

2) упругость растяжения (63):

$$D_2 = \sum_{i=1}^{p} \sum_{j=1}^{q-1} \left\| y^{ij} - y^{i,j+1} \right\|^2 + \sum_{i=1}^{p-1} \sum_{j=1}^{q} \left\| y^{ij} - y^{i+1,j} \right\|^2 \tag{63}$$

3) упругость изгиба (64):

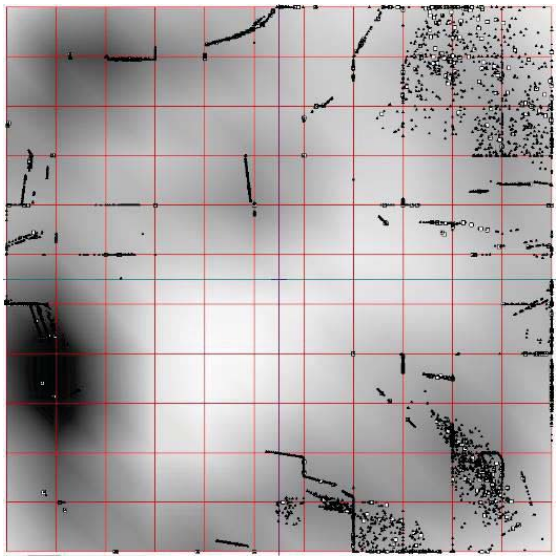
$$D_3 = \sum_{i=1}^{p} \sum_{j=2}^{q-1} \left\| 2y^{ij} - y^{i,j-1} - y^{i,j+1} \right\|^2 + \sum_{i=2}^{p-1} \sum_{j=1}^{q} \left\| 2y^{ij} - y^{i-1,j} - y^{i+1,j} \right\|^2 \tag{64}$$

Таким образом (65):

$$D = \frac{D_1}{|X|} + \lambda \frac{D_2}{pq} + \mu \frac{D_3}{pq} \to \min, \tag{65}$$

где $|X|$ – количество точек, λ, μ – коэффициенты «упругости».

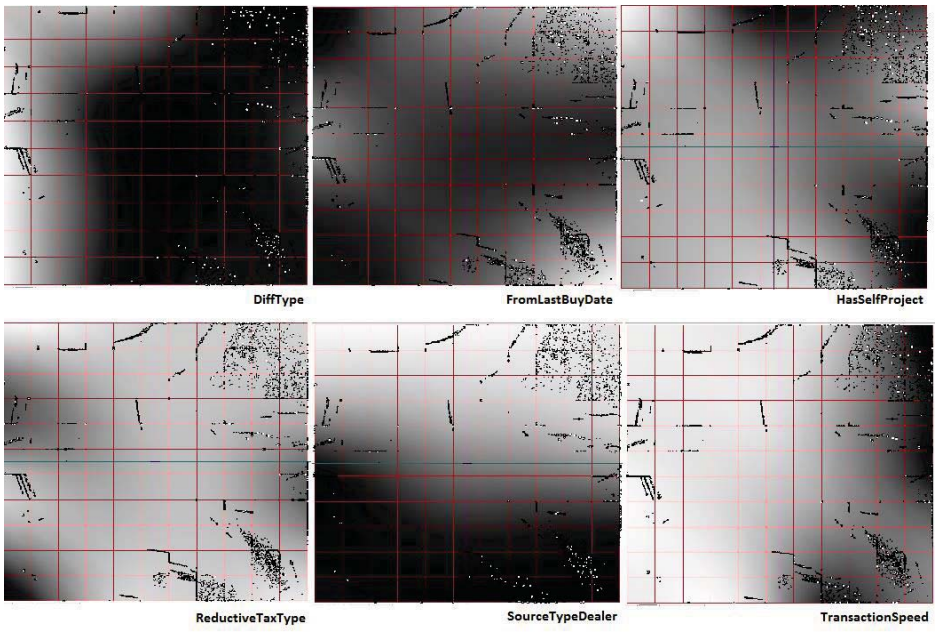
Выполним построение двумерной упругой карты с рекомендуемыми параметрами $\lambda = 0{,}01$, $\mu = 2$ и отобразим на плоскость, рисунок 27. Белые точки данных соответствуют удачным попыткам коммуникаций [34, С. 205].



Источник: составлено автором.

Рисунок 27 – Упругая карта. Распределение плотности данных

Как и в случае с картой Кохонена отобразим данные на упругой карте в разрезе информативных признаков. На рисунке 28 приведено 6 раскрасок по информативным признакам: 1) «Проявлял интерес к различным транспортным средствам»; 2) «Количество дней с последней покупки»; 3) «Самостоятельно интересовался покупкой»; 4) «Имеет упрощённую систему налогообложения; 5) «Обратился через дилера»; 6) «Скорость покупок».



Источник: составлено автором.

Рисунок 28 – Упругая карта в разрезе исследуемых признаков

Важным этапом подготовки данных является их нормализация. Для нормализации данных используются различные подходы. Самым распространённый подход – это стандартизация: центрирование и нормирование признаков (66) [17]:

$$\check{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_{x_i}}, \tag{66}$$

где $\check{x}_{ij}$ – стандартизованный признак x_i, σ_{x_i} – среднеквадратическое отклонение x_i, $\bar{x}_i$ – среднее значение.

Нормализацию можно выполнить иным способом: по эмпирической функции распределения (67):

$$X_i \to X_i' | x_{ij}' = x_{F(x_{ij})}, \tag{67}$$

где n – размер выборки, $x_{i1}, \dots, x_{in}$ – вектор уникальных значений признака X_i, $x_{i1}', \dots, x_{in}'$ – вектор уникальных значений признака X_i', $F(x_{ij})$ – эмпирическая функция распределения признака X_i (68):

$$F\left(x_{ij}\right) = \overline{P(x < x_{ıj})} = P\left(x < x_{ij}\right) + \frac{P\left(x = x_{ij}\right)}{2}, \tag{68}$$

$x_{F(x_{ij})}$ – является квантилем уровня $F\left(x_{ij}\right)$ для нормального распределения (69)

$$F(x) = N = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \tag{69}$$

Такая нормализация оказывается более подходящей для построения модели, но чтобы оценить вероятность отклика потребуется иметь всю таблицу эмпирической функции распределения. Целесообразно подобрать известную функцию распределения, максимально соответствующую эмпирической (70):

$$F\left(x_{ij}\right) - F_k\left(x_{ij}\right) \to min \tag{70}$$

Для описания распределения признаков «выручка» и «количество автомобилей в парке», подходит семейство гамма-распределений (71):

$$\Gamma(k,\theta) = \frac{\gamma(x,\frac{x}{\theta})}{\Gamma(k)\theta^k} \tag{71}$$

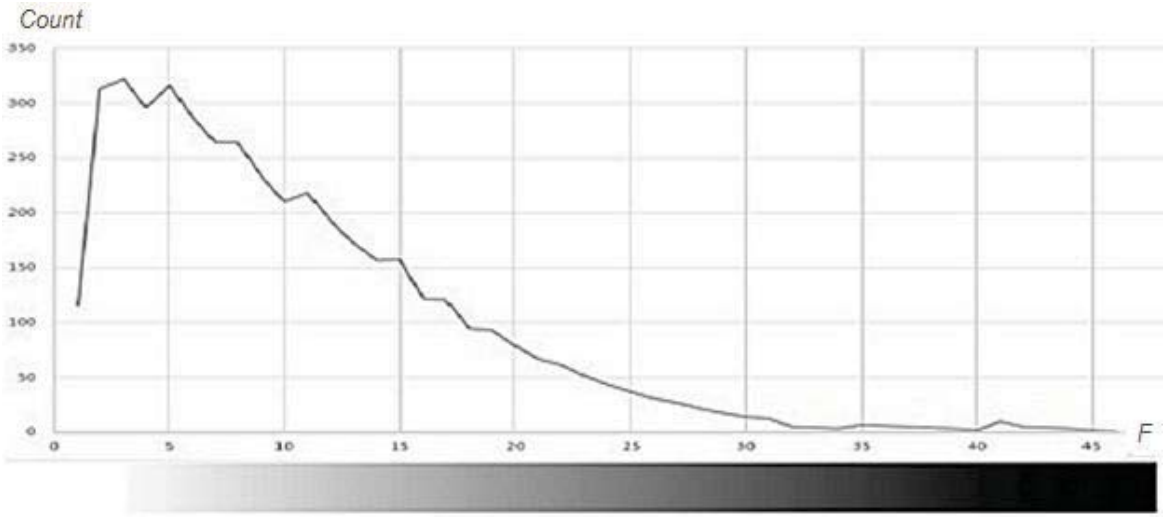
Требуется минимизировать (72)

$$F(x_{ij}) - \frac{\gamma(x_{ij},\frac{x_{ij}}{\theta})}{\Gamma(k)\theta^k} \to min \tag{72}$$

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

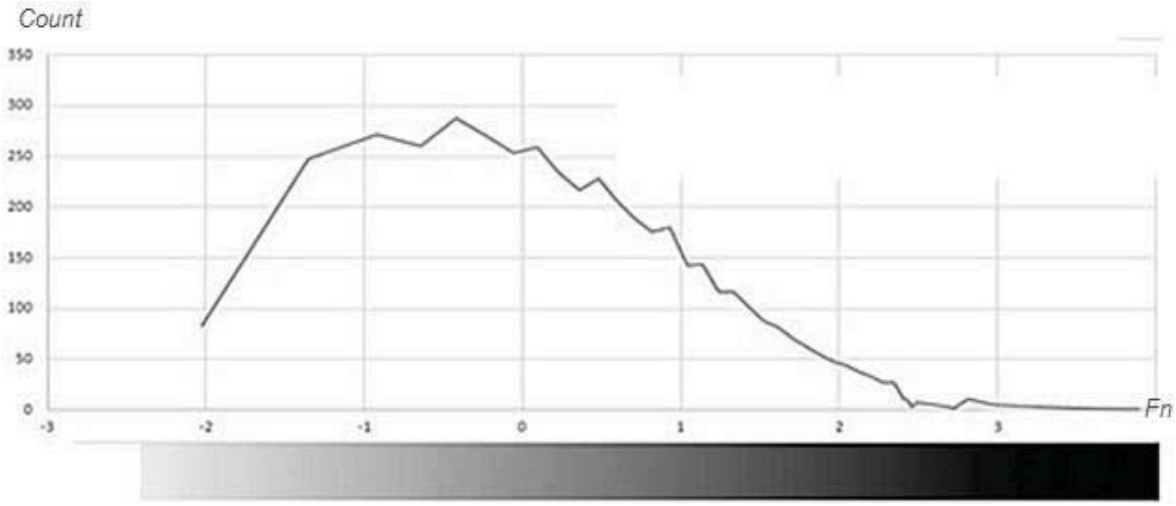
Пример такой нормализации проиллюстрирован на рисунке 29 и рисунке 30. На градиентной шкале более светлые участки соответствуют более низким значениям.

Нормализацию по эмпирической функции распределения полезно применять для отображения данных на различных поверхностях. Это помогает дать больший контраст между значениями признака там, где сосредоточена большая часть данных, рисунок 31.



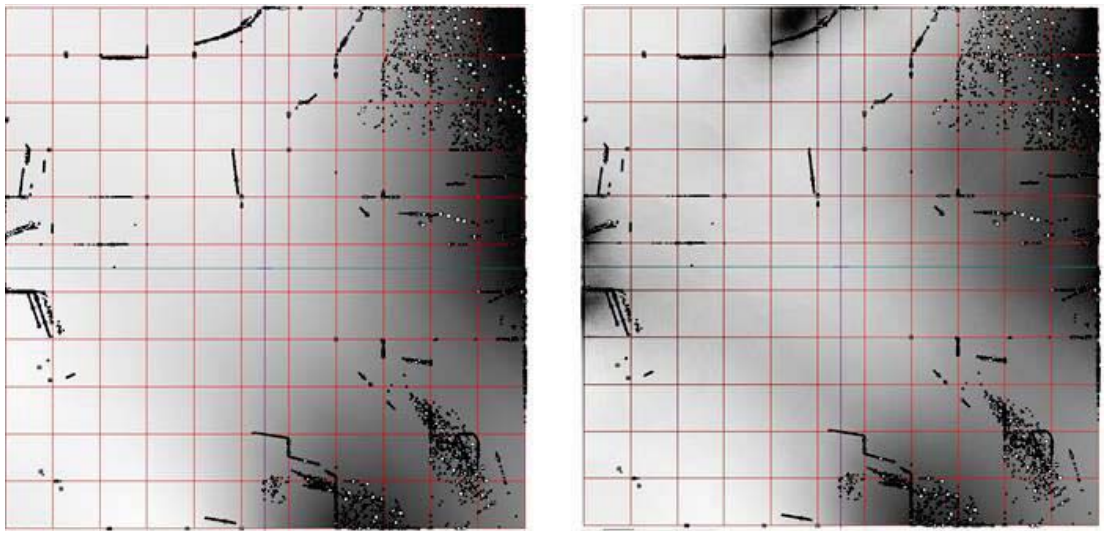
Источник: составлено автором.

Рисунок 29 – Распределение признака «Частота покупок» до нормализации



Источник: составлено автором.

Рисунок 30 – Распределение признака «Частота покупок» после нормализации

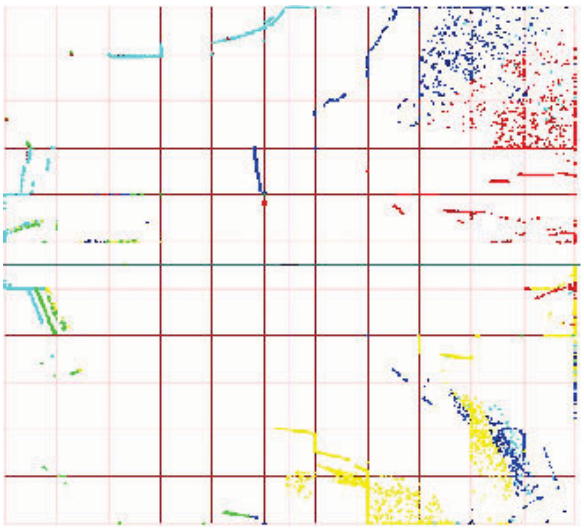


Источник: составлено автором.

Рисунок 31 – Распределение признака «Частота покупок» на упругой карте до нормализации (слева) и после нормализации (справа)

Нормализация по эмпирическому распределению даёт возможность не только корректировать оценку расстояния между объектами по отдельному признаку, но и уточнять расположение объектов в многомерном пространстве в целом. Это позволяет строить регрессионную модель, устойчивую к изменениям данных [34, 36].

Выполним кластеризацию методом k-means с 5 кластерами, изображёнными цветом на рисунок 32, и составим статистику, рисунок 33:



Источник: составлено автором.

Рисунок 32 – Кластеризация после построения упругой карты

Номер кластера	Avg Y	Count %	DiffType	FromLastBuyDate	HasSelfProject	ReductiveTaxType	InfoSourceTypeDealer	TransactionSpeed
1	8,26%	16,33%	0,974680617	-0,830773756	0,918211849	-0,329782911	-0,280614775	-0,176794014
2	5,52%	15,39%	-1,025882311	-0,701516086	0,483890702	-0,111857117	-0,553472892	-0,475491992
3	0,89%	39,31%	-0,446970002	0,602897661	-0,581769772	0,23913451	0,271289629	-0,672283852
4	5,54%	17,35%	0,611426113	-0,081218502	0,055098779	-0,08580076	0,891674421	1,294086541
5	5,64%	11,62%	0,588244376	0,178461624	-0,045523959	-0,069208282	-1,121381953	1,219527554

Источник: составлено автором.

Рисунок 33 – Статистика кластеров после применения упругой карты

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».

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».

Качество кластеризации можно оценить по соотношению средних внутрикластерных и межкластерных расстояний [16]. 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.

3.2 Построение регрессионной модели отклика клиентов на маркетинговые коммуникации

Увеличение эффективности использования данных о клиентах компании является стратегически важным управленческим инструментом. Это остаётся актуальным в кризисные периоды также и для лизинговых компаний. Развитый канал исходящих маркетинговых коммуникаций с опорой на клиентскую базу является потенциально полезным внутренним ресурсом лизинговой компании. Эффективность маркетинговых коммуникаций можно увеличить за счёт комплекса мер по улучшению маркетинга, подготовки продающих кадров, совершенствования математического и информационного обеспечения. В данном разделе приведено описание средств и методов математического анализа и моделирования, повышающих качество утилизации клиентской базы, и, как следствие, эффективность маркетинговых коммуникаций [36].

Как было отмечено выше, за численное значение эффективности маркетинговых коммуникаций, как правило, берётся конверсия, как

отношение количества коммуникаций с целевым действием ко всем коммуникациям (73):

$$Conv = \frac{TS}{TS + TF}, \quad (73)$$

где $Conv$ – конверсия, TS – количество успешных попыток коммуникаций, TF – количество неудачных попыток коммуникации.

Также не исключён выбор иных показателей эффективности, таких как ROI – окупаемость инвестиций (return on investment).

Конверсия выбрана в качестве оптимизируемого показателя маркетинговых коммуникаций, поскольку объём данных о положительных исходах, приводящих к окупаемости, по сравнению с объёмом данных об общем числе попыток (0.6% при оптимизации конверсии, и порядка 0,05% без неё) незначителен, другими словами, мы имеем несбалансированность выборки. Использование показателя ROI данной ситуации приведёт к неустойчивости построенной модели.

Поскольку ресурсы каналов маркетинговых коммуникаций ограничены, следует вначале искать контакта с потенциально высококонверсионными клиентами. Это приводит к необходимости ранжирования клиентов в порядке убывания ожидаемой вероятности отклика. Оптимизируемый показатель будет вычисляться по формуле (74):

$$\varphi = \sum_{i=1}^{n} (y_i - \hat{P}(y_i = 1))^2 \to min, \quad (74)$$

где n – количество клиентов, y_i – исход коммуникации с i-м клиентом (1, если успешно, 0 – в обратном случае), $\hat{P}(y_i = 1)$ – оценка для вероятности успеха коммуникации с i-м клиентом.

Подготовка данных является неизбежным шагом перед непосредственно моделированием. Она заключается в произведении их очистки, и, при необходимости, решении проблемы пропусков значений. Используются разные подходы к решению этой задачи [20, 27]:

- метод удаления неполных векторов (casewise deletion);

- метод замены средними (mean substitution);
- метод замены условными средними значениями (imputation by regression);
- метод наполнения выборочными значениями (hot deck imputation);
- метод максимального правдоподобия (Expectation-Maximization, EM algorithm).

По результатам исследований [24] в большинстве методов анализа наиболее устойчивый результат даёт метод заполнения выборочными значениями (hot deck imputation) при наличии небольшого числа пропущенных значений. Метод состоит в том, чтобы для каждого элемента (x_{01}, x_{02}) с пропуском найти два ближайших элемента (x_{11}, x_{12}), (x_{21}, x_{22}), чей искомый признак известен. После чего производится линейная интерполяция для нахождения признака x_{02} (75) [20, С. 75]:

$$x_{02} = x_{12} + \frac{x_{22} - x_{12}}{x_{21} - x_{11}}(x_{01} - x_{11}) \quad (75)$$

Исходя из изложенных выше соображений, для заполнения пропущенных значений был использован этот метод.

Специфика задачи оптимизации маркетинговых коммуникаций состоит в сортировке задач менеджера прямых продаж по убыванию вероятности отклика клиента. Таким образом, необходимо обеспечить непрерывность оценки этой вероятности. Эта задача хорошо решается с использованием логистической регрессии.

В решении поставленной задачи целесообразно использовать логистическую регрессию (76):

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}, \quad (76)$$

где $g(\theta^T x)$ – логистическая (или сигмоидная) функция, θ – вектор признаков x.

Для построения модели имеет смысл использовать различные признаки объектов. Ими могут являться как готовые признаки в клиентской базе, так и синтетические, построенные по экспертным рекомендациям. Вместе с тем необходимо учитывать, что в маркетинговых коммуникациях наибольший вклад в оценку вероятности покупки вносят такие характеристики, как давность покупки (Recency), их частота (Frequency) и количество потраченных денег (Monetary). Это наблюдение является основой современного RFM-анализа [152, 118]. Виды зависимостей вероятности покупки от прошедшего времени, частоты и количества потраченных денег рассмотрены в разделе 1.2.

В рассмотренных исследованиях приводится анализ маркетинговых коммуникаций с физическими лицами. Модель зависимости вероятности отклика на маркетинговые коммуникации бизнес-клиентов может иметь несколько иной вид.

В предыдущем разделе приведён анализ дифференцирующей способности признаков, а также рассмотрены нелинейные методы, предназначенные для визуального анализа признаков.

Для построения логистической модели был использован программный продукт R (это свободно распространяемая в рамках проекта GNU программа с открытым исходным кодом для статистических вычислений). Результат работы приведён на рисунке 34.

Z-value (Z-критерий Фишера) – показатель отклонения случайной величины от её математического ожидания по отношению к стандартной ошибке (77):

$$Z = \frac{\bar{X} - M[X]}{\sigma}, \tag{77}$$

где $\bar{X}$ – выборочное среднее, $M[X]$ – математическое ожидание, σ – стандартная ошибка.

При проверке гипотезы на равенство переменной нулю, полагают (78)

$$H_0: M[X] = 0 \quad (78)$$

Тогда, чем больше по модулю показатель Z, тем меньше вероятность Pr(>|Z|) того, что математическое ожидание случайной величины равно нулю.

```
Call:
glm(formula = Y ~ FromLastTransaction + TransactionCount + DiffType +
HasSelfProject + InfoSourceTypeMOP + ReductiveTaxType + Div2 +
Segment1 + Segment2 + Segment3 + Segment4 + Segment6 + CountTC,
family = binomial("logit"), data = check)

Deviance Residuals:
Min        1Q        Median      3Q         Max
-3.5326    -0.2927   -0.1500     -0.0427    4.0263

Coefficients:
                      Estimate    Std. Error  Z-value   Pr(>|Z|)
(Intercept)           -3.3406486  0.1751501   -19.073   < 2e-16   ***
FromLastTransaction   -0.0153486  0.0008070   -19.019   < 2e-16   ***
TransactionCount       0.0790810  0.0094304   8.386     < 2e-16   ***
DiffType               0.3697491  0.0889078   4.159     3.20e-05  ***
HasSelfProject         0.6879864  0.0898078   7.661     1.85e-14  ***
InfoSourceTypeMOP      0.1833434  0.0818873   2.239     0.025158  *
ReductiveTaxType      -0.1853607  0.0966481   -1.918    0.055125  .
Div2                   0.2350859  0.0877988   2.678     0.007416  **
Segment1               0.6652471  0.1473257   4.515     6.32e-06  ***
Segment2               0.7127300  0.2099243   3.395     0.000686  ***
Segment3               0.8748192  0.3146928   2.780     0.005437  **
Segment4               0.6459596  0.1689045   3.824     0.000131  ***
Segment6               0.4657387  0.1841801   2.529     0.011448  *
CountTC                0.0012856  0.0004383   2.933     0.003353  **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Источник: составлено автором.

Рисунок 34 – Логистическая регрессия вероятности отклика на маркетинговые коммуникации в R

По результатам построения логистической регрессии наиболее значимыми получились следующие признаки: «Дней с последней покупки», «Количество покупок», «Клиент интересовался разными транспортными средствами».

Качество логистической модели, как классификатора, можно измерить вычислив площадь под графиком ROC-кривой (AUC – area under curve). Эта кривая отображает соотношение специфичности (Specificity) и чувствительности (Sensitivity) модели в зависимости от порога отсечения (79, 80).

$$Sensitivity\ (TPR) = \frac{TP}{P} = \frac{TP}{TP + FN}, \quad (79)$$

$$Specificity\ (SPC) = \frac{TN}{N} = \frac{TN}{FP + TN}, \quad (80)$$

где TP – правильно выявленные положительные исходы, TN – правильно выявленные отрицательные исходы, FP – неправильно выявленные положительные исходы, FN – неправильно выявленные отрицательные исходы.

Факторы модели приведены в таблице 19.

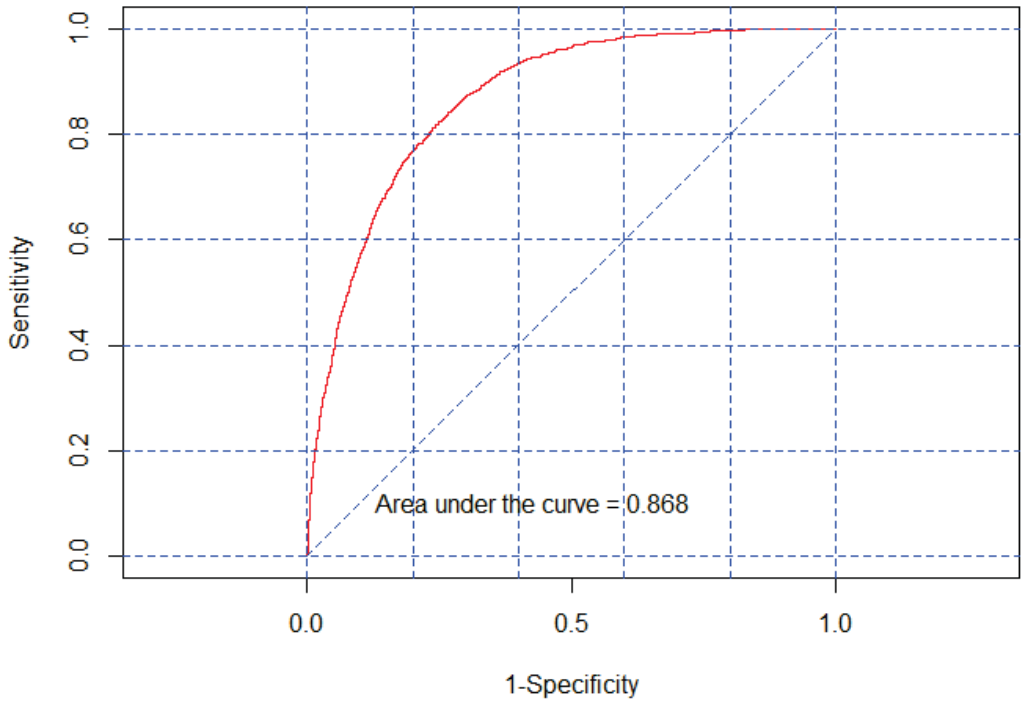
Таблица 19 – Русскоязычная интерпретация факторов модели

Признак	**Интерпретация**
FromLastTransaction	Дней с последней покупки
TransactionCount	Количество покупок
DiffType	Интересовался разным типом транспортных средств
HasSelfProject	Наличие самостоятельного обращения клиента
InfoSourceTypeMOP	Клиента привёл менеджер отдела продаж
ReductiveTaxType	Упрощённая система налогообложения
Div2	Группа офисов продаж 2 (по географическому признаку)
Segment1	Интересовался крупнотоннажным коммерческим транспортом
Segment2	Интересовался среднетоннажным коммерческим транспортом
Segment3	Интересовался лёгким коммерческим транспортом
Segment4	Интересовался легковыми автомобилями
Segment6	Интересовался оборудованием
CountTC	Количество автомобилей в парке

Источник: составлено автором.

При наихудшем сценарии (случайная классификация) значение показателя AUC равно 0.5. Значение AUC, равное 0,87 можно считать

признаком хорошей классификации, рисунок 35. Например, при пороге отсечения, дающем 15% вероятности отклика клиентов на маркетинговые коммуникации, в сегменте, отсечённом моделью, вероятность отклика не превышает 1.5%.



Источник: составлено автором.

Рисунок 35 – ROC-кривая для построенной модели отклика на маркетинговые коммуникации на тестовых данных

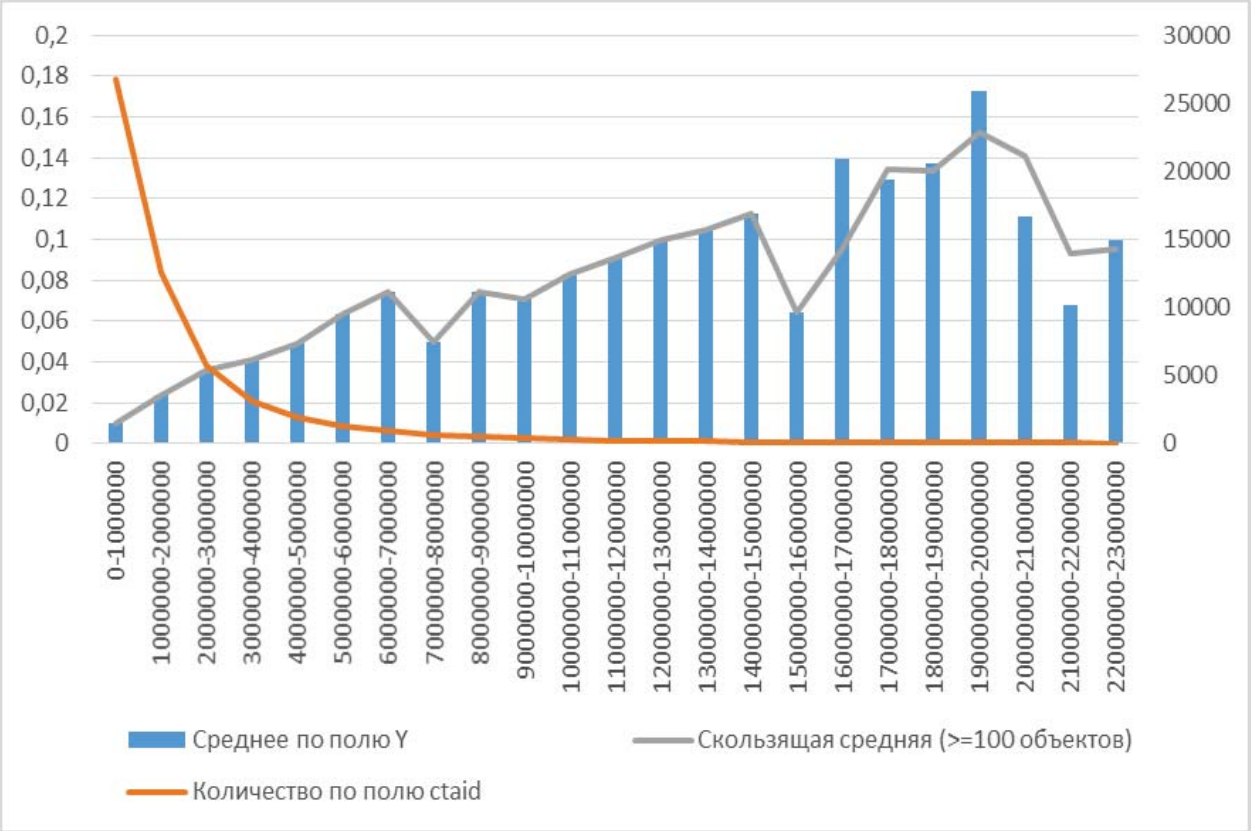
С учётом специфики экономической задачи признак «Количество покупок» оказался несколько более значимым, чем «Количество затраченных денег», таблица 20.

Таблица 20 – Z-статистика для признаков «Количество покупок» и «Количество затраченных денег» в моделях

Признак	**Z-значение**
TransactionCount	8.386
TransactedPrice	5.470

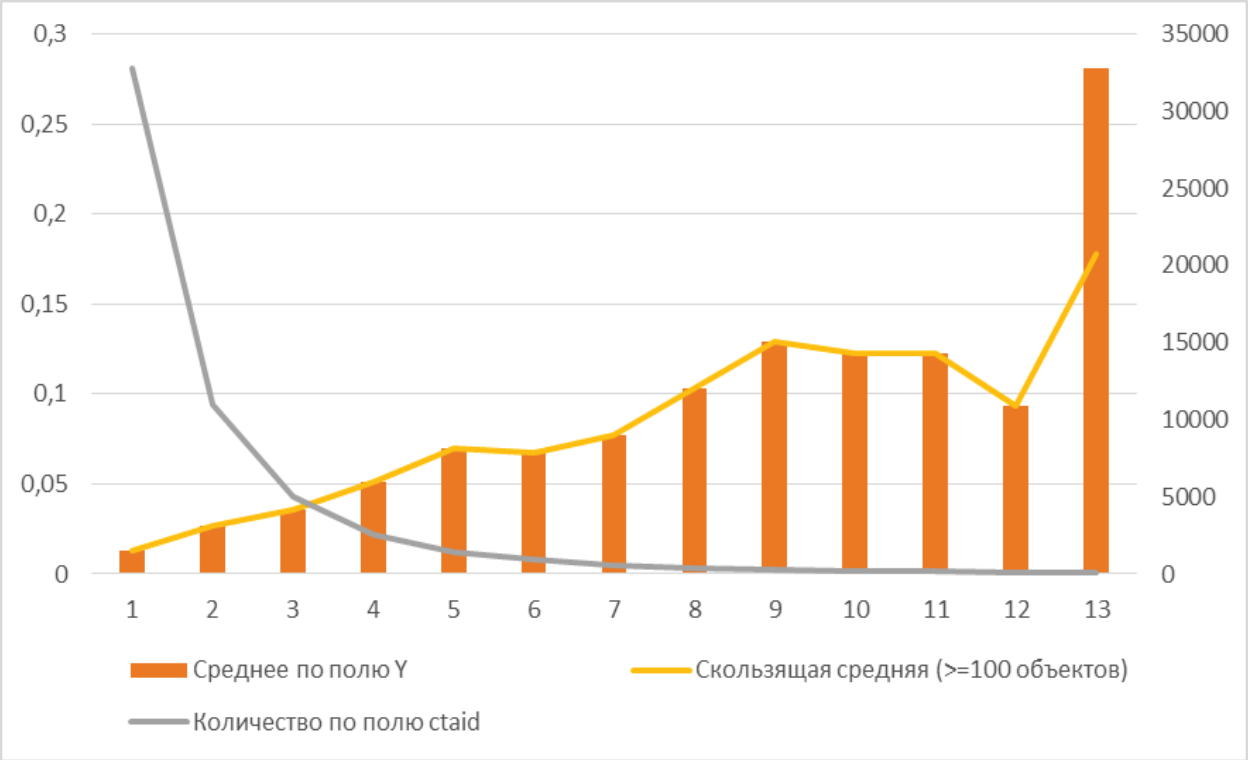
Источник: составлено автором.

На рисунке 36 и рисунке 37 проиллюстрирована зависимость отклика от суммы затраченных денег и количества покупок соответственно.



Источник: составлено автором.

Рисунок 36 – Зависимость вероятности отклика от суммы затраченных денег



Источник: составлено автором.

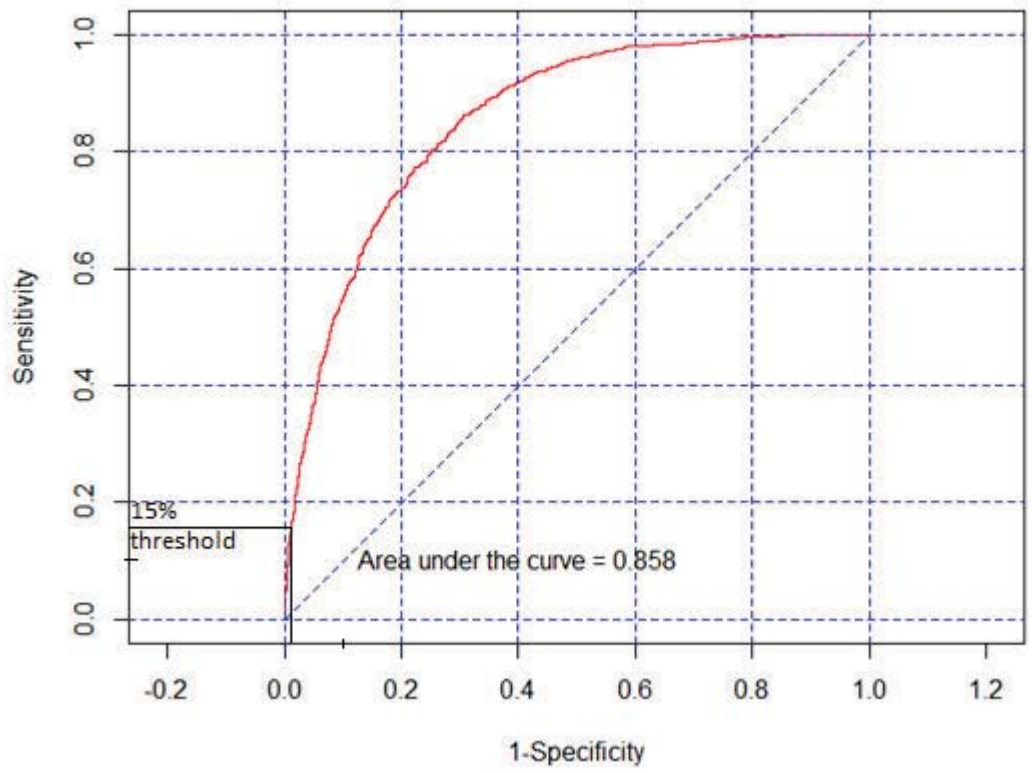
Рисунок 37 – Зависимость вероятности отклика от количества покупок

Тем не менее, эти признаки имеют очень высокую корреляцию: 0,807 в выборке и 0,889 в генеральной совокупности. Поэтому, автор счёл целесообразным включить в модель один из этих признаков, таблица 21, рисунок 38.

Таблица 21 – Матрица корреляции признаков «Количество покупок», «Количество затраченных денег» и прогнозируемой переменной

	TransactedPrice	TransactionCount
Y	0,12462	0,1622034
TransactedPrice	1	0,8071239

Источник: составлено автором.



Источник: составлено автором.

Рисунок 38 – ROC-кривая для построенной модели отклика на маркетинговые коммуникации на тестовых данных с использованием признака «Количество затраченных денег»

Построенная и описанная выше модель используется для ранжирования клиентов в маркетинговых коммуникациях.

Отклик клиентов на маркетинговые коммуникации хорошо описывается математической моделью. Построенная методика использования клиентской базы позволяет кратно повысить эффективность маркетинговых коммуникаций с помощью средств математического моделирования, современных методов анализа, очистки и визуализации данных, а также численных методов оценки распределения признаков.

ГЛАВА 4
АВТОМАТИЗАЦИЯ ПРОЦЕССОВ ОБРАБОТКИ ДАННЫХ И УПРАВЛЕНИЯ МАРКЕТИНГОВЫМИ КОММУНИКАЦИЯМИ

4.1 Разработка архитектуры объединённого хранилища данных для маркетинговых коммуникаций

Для хранения результатов структуризации и очистки данных, а также для эффективного управления маркетинговыми коммуникациями, необходимо развернуть хранилище данных, на базе которого будут запущены автоматизированные процессы.

Потребность в объединённом хранилище данных также возникает в виду появления новой непротиворечивой структуры данных, вбирающей в себя полезную с точки зрения маркетинговых коммуникаций информацию из всех систем [38].

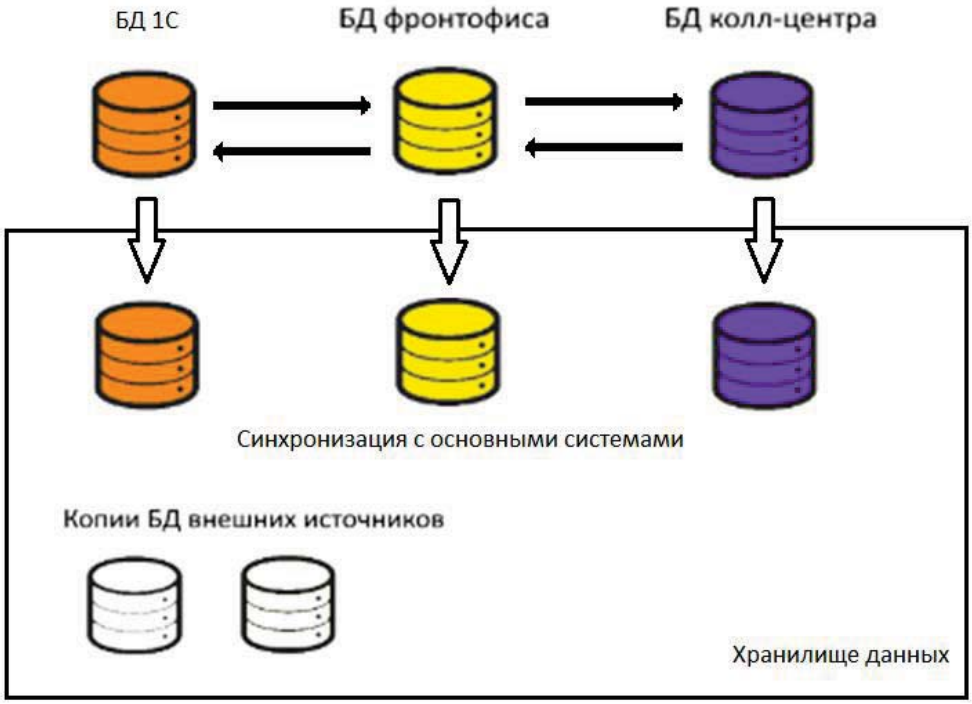
Исходные данные для автоматизации содержатся в базах данных следующих систем:

- система автоматизированной обработки звонков;
- система автоматизированного фронт-офиса;
- 1С.

Кроме перечисленных систем имеются данные из других источников, таких, как 2Gis, auto.ru и прочие источники, ценные с точки зрения содержания полезной информации о юридических лицах, например, такой, как, данные об автопарке. Данные из внешних источников зачастую являются слабоструктурированными и слишком загрязнены.

Объединённое хранилище данных является хранилищем второго уровня, так как его внедрение происходит опосредовано, через уже существующее хранилище необработанных данных.

Схема потоков данных между базами данных основных систем представлена на рисунке 39.



Источник: составлено автором.

Рисунок 39 – Схема потоков данных между основными системами

С точки зрения непротиворечивости данных о клиентах системы фронт-офиса и колл-центра хорошо интегрированы, однако, имеются серьёзные проблемы с синхронизацией других сущностей, например, таких, как «договор лизинга».

Обратная ситуация наблюдается в интеграции систем фронт-офиса и 1С. Данные, касающиеся договоров лизинга хорошо синхронизированы, в то время, как в одной из систем несколько договоров лизинга могут относиться к одному юридическому лицу, а в другой – к нескольким.

Данные систем БД колл-центра и 1С вообще не синхронизированы между собой.

Задача, решаемая в рамках проектирования объединённого хранилища данных состоит в сборе информации из всех перечисленных систем в

непротиворечивую структуру, а также, в обогащении информации о клиентах за счёт баз данных, полученных из внешних источников.

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.

Укрупнённая архитектура программного комплекса представлена на рисунке 40.

Объединённое хранилище данных развёрнуто на выделенном сервере. Данные загружаются на сервер непосредственно из хранилища, после чего вся обработка данных осуществляется за счёт ресурсов сервера.

Основные элементы архитектуры базы данных программного комплекса представлены на рисунке 41.

Разработанный и внедрённый программно-аппаратный комплекс состоит из модулей, предназначенных для решения частных задач.

Модуль нечёткого сравнения идентификационных данных:

- сбор идентификационных данных контрагентов из разных систем;
- нечёткий поиск дубликатов по набору различных характеристик;
- перенос данных промежуточной структуры (карты идентификаторов) на выделенный сервер.

Модуль сбора, очистки и стандартизации данных:

- очистка, стандартизация и перенос первичных характеристических данных на выделенный сервер с сохранением структуры.

Модуль преобразования характеристических данных:

- вычисление вторичных характеристических данных контрагентов, используемых для вычисления вероятности по модели.

Модуль объединения данных:

- вычисление характеристик групп контрагентов на основе карт идентификаторов и сохранение в новой структуре данных.

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

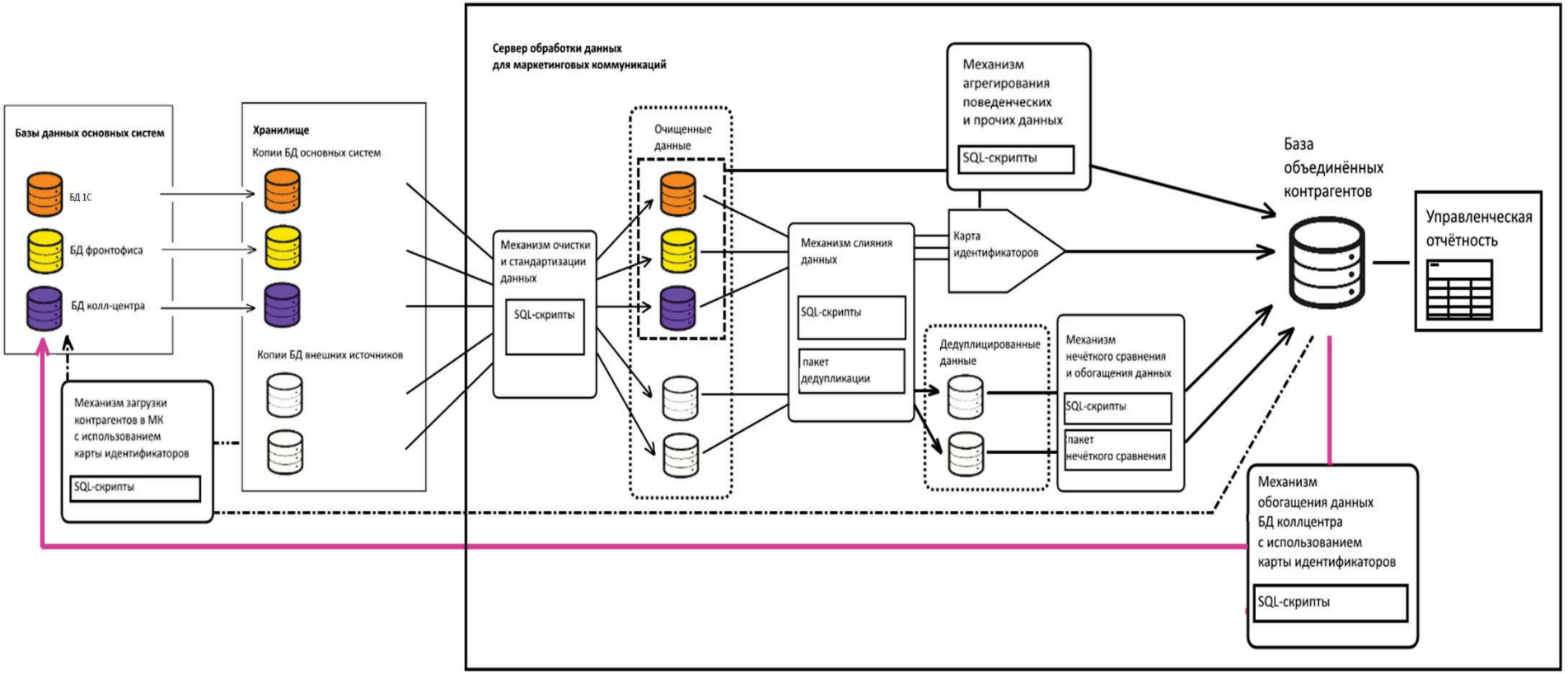
- формирование выборки контрагентов в соответствии с заданными критериями МК;
- отсечение клиентов с низким потенциалом отклика, рассчитанным по логистической модели.

Модуль загрузки контрагентов для коммуникации в рамках МК:

- загрузка списка контрагентов для обработки в рамках МК.

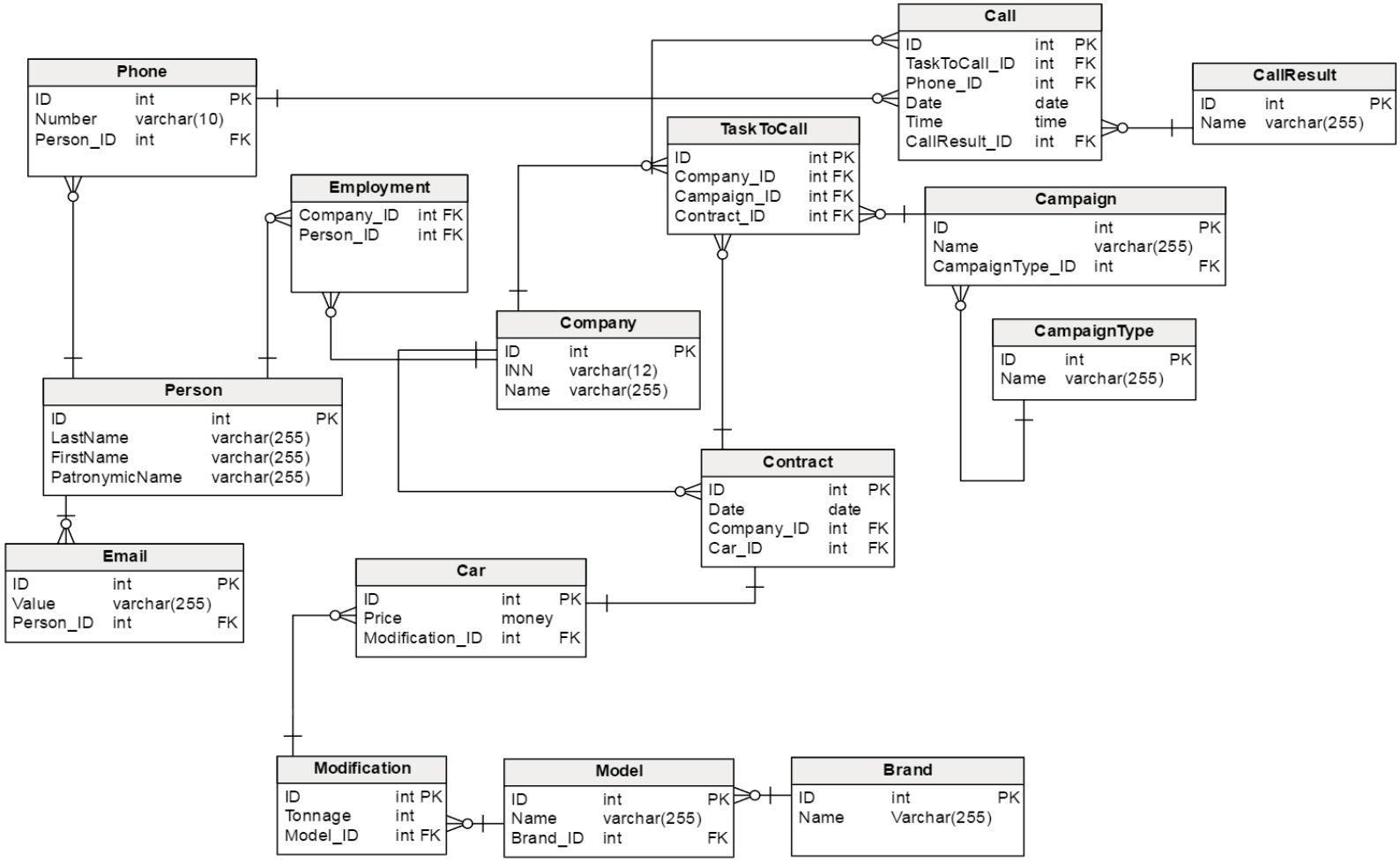
Модуль нечёткого сравнения является отторгаемым и может быть внедрён в другие системы.

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.



Источник: составлено автором.

Рисунок 40 – Схема разработанного программного комплекса



Источник: составлено автором.

Рисунок 41 – Структура базы данных

4.2 Автоматизация процессов структуризации, очистки и агрегирования слабоструктурированных данных

Работа с данными о клиентах условно разделяется на две составляющие: работа с данными из внутренних систем, и работа с данными из внешних источников. Упрощённо автоматизированный процесс работы сервера объединённого хранилища данных можно представить в виде набора последовательных шагов. Варианты шагов могут различаться в зависимости от периодичности выполнения. На некоторых шагах процедуры выполняются параллельно (например, импорт данных).

Работа с данными о клиентах из внутренних систем:

1. очистка и импорт данных из основных систем;
2. подготовка таблиц для нечёткого поиска дубликатов сущностей клиентов;
3. поиск групп дубликатов сущностей клиентов для новых записей по модифицированному составному ключу (в короткий период) или полноценная дедупликация (в длинный период);
4. генерация ключей для новых сущностей клиентов, полученных в результате дедупликации. Создание таблицы-отношения старых и новых ключей (карты ключей);
5. выбор мастер-сущностей клиентов для экспорта данных в основные системы;
6. обновление базы объединённых дубликатов клиентов;
7. ассоциирование контактных данных и характеристик с дедуплицированными контрагентами, полученных с помощью карты ключей, сохранение контактных данных в базе объединённых дубликатов клиентов;
8. импорт характеристик контрагентов из внешних источников;
9. обновление базы данных колл-центра для мастер-сущностей в части контактных данных и полученных характеристик.

Очистка и проверка на корректность телефонов, e-mail, фамилий, имён и отчеств контактных лиц контрагентов осуществляется с помощью регулярных выражений [37], таблица 22, таблица 23, таблица 24.

Таблица 22 – Пример очистки телефонов

Исходный	**Очищенный**
+7 (351) 2664386	3512664386
9178567854#6	9178567854
4956518179#2203	4956518179

Источник: составлено автором.

Таблица 23 – Пример очистки e-mail

Исходный	**Очищенный**
[rekardooffice]@vinf.ru	rekardooffice@vinf.ru
<jlebedeva>@ppe.ru	jlebedeva@ppe.ru
istok-m@mail.ru	istok-m@mail.ru

Источник: составлено автором.

Таблица 24 – Пример дедупликации

Входной ключ	**Выходной ключ**	**Safo Id**	**1C Id**	**Capella Id**	**Наименование**	**ИНН**	**Адрес**
114581	941459	515	-	-	ФЛ «Дальэнергомонтаж»	2702010860	68013, город Хабаросвк, улица Ленина, д.10
116831	941459	-	-	0x9CE14	ФЛ «Дальэнергомонтаж»	-	680013, Хабаровский край, г. Хабаровск, ул. Ленина, 10
941459	941459	-	-	0x6574D	фл “Дальэнергомонтаж”	-	Хабаровск, улица Ленина, 10
100353	941459	-	-	0x07E46	Дальэнергомонтаж	2702010860	-
102564	941459	-	-	0x6B538	Дальэнергомонтаж, фл	2702010860	ул. Ленина, 10
122852	941459	-	-	0x78D65	Дальэнергомонтаж	2702010860	-

Источник: составлено автором.

Для очистки и проверки на корректность ФИО, телефонов и адресов использованы регулярные выражения на базе подгружаемых CLR-функций.

Пример регулярного выражения для очистки ФИО:

'(?i)^[^а-яё]+|[^а-я- ё]+|[^а-яё]+$'

Пример регулярного выражения для проверки на корректность e-mail:

'(?i)@\d*\.|(...).?\1.?\1|^(.)\2*@|^net@|mail@mail|^([^.@][^.@])@\3\.[^.@]+$|^(?!.*?\S{8})|^(?![^@]{2,}@[^.@]{2,}\.[^.@]{2,})'

Пример регулярного выражения для очистки телефона

'123456|(.)\1{5}|(..)\2{3}|^800|^(?!\d{10}$)'

В некоторых случаях алгоритм дедупликации не способен распознать различия между сущностями клиентов в силу объективных причин. В частности, если известен регион и наименование контрагента, то совпадение этих двух параметров ещё не говорит о совпадении сущностей. Для разрешения этого вопроса можно прибегнуть к дополнительному алгоритму, например, последовательно сравнивая телефоны и e-mail: в случае их совпадения сущности контрагентов с большой вероятностью окажутся дубликатами. Однако, в случае наличия большого количества ложной информации в базе данных, а именно, аномально часто встречающихся телефонов и e-mail, можно совершить ошибку, объединив эти сущности. Поэтому следует учитывать этот фактор при уточнении результатов дедупликации.

Для выбора мастер-сущностей был использован алгоритм, суть которого заключается в выборе сущности с максимально высокой заполненностью полей. Общая заполненность линейно зависит от заполненности каждого поля. В случае небольшого количества (до 30) полей возможно задать модель подсчёта заполненности, в которой заполненность поля с наибольшим коэффициентом будет определяющей.

Работа с данными о клиентах из внешних источников:

1. загрузка данных, полученных из других систем (для синтаксического разбора баз данных сайтов используется программный пакет на языке C#);
2. очистка и импорт данных;
3. дедупликация;
4. структуризация;
5. построение агрегированных характеристик;
6. загрузка новых сущностей в базу данных колл-центра (опционально).

Техническая реализация очистки данных и дедупликации в работе с данными из внешних источников различается незначительно.

4.3 Автоматизация управления маркетинговыми коммуникациями на основе регрессионной модели

В системе автоматизации колл-центра основной сущностью является задача на звонок клиенту. Соответственно, управление исходящими маркетинговыми коммуникациями сводится к менеджменту таких задач. Каждая задача обрабатывается в рамках одной из множества маркетинговых кампаний, которые отражают уникальные торговые предложения.

К моменту постановки задачи на звонок вся необходимая информация уже находится в базе данных системы.

Каждая маркетинговая кампания нацелена на определённую аудиторию, сегмент клиентской базы.

Определение сегмента контрагента производится либо на основе экспертных критериев, либо на основе полученных кластеров. Часто эти методы применяются совместно.

После выбора клиентов из нужного сегмента, отбираются те, которым разрешается звонить по контактной политике компании.

На языке запросов T-SQL это выглядит примерно следующим образом: рисунок 42.

```
declare @Threshold float = 0.15;
with Model as(
	select
		id,
		exp(value)/(1+exp(value)) prob
	from
		#Target /*Таблица клиентов из целевого сегмента*/
		outer apply(
			select
				-3.3406486
				+ FromLastTransaction * -0.8748192
				+ TransactionCount *	0.7127300
				+ DiffType *		0.6652471
				+ HasSelfProject *	0.6879864
				+ InfoSourceTypeMOP * 0.6459596
				+ ReductiveTaxType *	-0.1853607
				+ Div2 *			0.2350859
				+ Segment1 *			0.3697491
				+ Segment2 *			0.0790810
				+ Segment3 *			0.0153486
				+ Segment4 *			0.1833434
				+ Segment6 *			0.4657387
				+ CountTC *		0.0012856
				value
			) Calc
	)
select *
from(
	select
		id,
		prob,
		avg(prob) over(order by prob desc) avgProb
	from Model
	) T
where T.avgProb > @Threshold
```

Источник: составлено автором.

Рисунок 42 – Расчёт вероятности отклика на T-SQL

Для каждой сущности рассчитывается значение вероятности отклика по построенной регрессионной модели. После чего создаются задачи по наиболее конверсионным по оценке модели клиентам с заданным порогом отсечения по вероятности [36, С. 168].

В общем случае сегменты контрагентов могут пересекаться, поэтому необходимо задать приоритет определения маркетинговой кампании для созданных задач. После создания задачи передаются оператору колл-центра системой распределения задач.

ГЛАВА 5
ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ И ОЦЕНКА ЭКОНОМИЧЕСКОГО ЭФФЕКТА ОПТИМИЗАЦИИ МАРКЕТИНГОВЫХ КОММУНИКАЦИЙ

Разработанный программный комплекс позволил повысить качество данных о клиентах компании [37]. Основные показатели очистки данных проиллюстрированы в таблице 26, таблице 27 и на рисунке 46.

Таблица 26 – Краткая статистика по очистке телефонов

Наименование параметра	Значение
Всего очищенно телефонов	3384187
Из них уникальных	1846700 (сжатие 1,83)
Признаны корректными	1817812 (0,98)
Признаны некорректными	28888 (0,02)

Источник: составлено автором.

Таблица 27 – Краткая статистика по очистке email

Наименование параметра	Значение
Всего очищенно e-mail	955166
Из них уникальных	349030 (сжатие 2,70)
признаны корректными	345357 (0,99)
признаны некорректными	3673 (0,01)

Источник: составлено автором.

Выявлены основные характеристики, влияющие на результат маркетинговых коммуникаций:

- «Прошло времени с последней покупки»
- «Количество покупок»
- «Интересовался ли разными транспортными средствами»
- «Впервые обратился через менеджера компании»

Характер влияния этих характеристик отображён с помощью простых диаграмм, а также с помощью упругих карт. Полученные знания

способствуют снижению неопределённости и принятию верных управленческих решений.

На рисунке 49 изображены конечные потребители разработанного комплекса:

1. Аналитик получает таблицы в новой схеме данных. С помощью них он имеет возможность писать скрипты для управления потоком контрагентов-участников МК, а также формировать отчёты, не заботясь о том, как и откуда получить правильные данные.

2. Колл-центр получает качественный поток контрагентов-участников МК с потенциально высокой конверсией и полноценным соблюдением контактной политики.

3. Управленческий аппарат получает качественную управленческую отчётность, а также, возможность оперативно сформировать новый вид отчётности.

4. Отдел маркетинга получает отчёт о результатах маркетинговых коммуникаций.

Выявлен комплекс факторов, влияющий на конверсию маркетинговых коммуникаций в целом:

- качество контактных данных;
- наличие дубликатов;
- актуальность данных;
- качество скриптов;
- структура клиентской базы;
- профессионализм менеджера прямых продаж;
- точность математической модели;
- работа менеджера отдела продаж.

Выявлены: динамика создания дубликатов в базе данных и устранены её последствия, рисунок 43, рисунок 46; зависимость конверсии маркетинговых коммуникаций от стажа менеджера прямых продаж, рисунок 44, от неповеденческих признаков клиентов, рисунок 47.