содержание

[введение 2](#_Toc135433325)

[1 теоретическая часть 3](#_Toc135433326)

[1.1 Основные понятие из статистики 3](#_Toc135433327)

[1.2 Классификация 8](#_Toc135433328)

[2 практическая часть 13](#_Toc135433329)

[2.1 Предобработка данных 13](#_Toc135433330)

[2.1 Прогнозирование с помощью логистической регрессии 16](#_Toc135433331)

[Заключение 20](#_Toc135433332)

[список использованных источников 21](#_Toc135433333)

введение

В современном мире бизнес соперничает в высокотехнологичной арене. Компании всех масштабов применяют новейшие технологии, чтобы повысить доходы и улучшить свою конкурентоспособность. Сегодня, большинство компаний осознает, что данные — это главный инструмент увеличения прибыли и привлечения, и удержания клиентов. Именно в свете этого и возникает необходимость использования клиентской аналитики, для понимания потребностей и поведения клиентов.

Клиентская аналитика — это процесс сбора и анализа данных обо всех клиентах организации. Обычно, это включает в себя информацию о покупках, обращениях в службу поддержки, использованную информацию на веб-сайте и многое другое. Данный тип аналитики помогает организациям понимать, какие продукты и услуги наиболее востребованы, а также локализовать и решать возникающие проблемы.

Таким образом, клиентская аналитика играет огромную роль в современном бизнесе. Она позволяет организациям повысить эффективность своих процессов, удерживать клиентов и повышать доходы.

Цель данной курсовой работы — клиентский анализ, прогнозирование оттока клиентов магазина.

Задачи, которые решаются в рамках данной курсовой работы:

* изучение научной и методической литературы по математической статистике;
* использование знаний из области математической статистики с использованием современных средств обработки данных: языка программирования R;
* построение статистической модели и ее оценка;
* прогнозирование с помощью логистической регрессии;
* изучить оформление документации.

1 теоретическая часть

1.1 Основные понятие из статистики

В современном обществе важную роль в механизме управления экономикой выполняет статистика. Независимо от уровня и стадии экономического развития, статистика на протяжении сотен лет своего существования всегда выступала как необходимый и эффективный инструмент государственного управления и одновременно как наука, исследующая количественную сторону массовых явлений.

Особенность статистики заключается в том, что статистические данные сообщаются в количественной форме, т. е. статистика говорит языком цифр, отображающих общественную жизнь во всем многообразии ее проявлений.

Выполняя самые разнообразные функции сбора, систематизации и анализа сведений, характеризующих экономическое и социальное развитие общества, она всегда играла роль главного поставщика факторов для управленческих, научно-исследовательских и прикладных практических нужд различного рода структур, организаций и населения. [1.1, с. 5]

В настоящее время под статистикой понимается:

* совокупность итоговых сведений, количественно характеризующих различные стороны общественной жизни: производство, распределение и обмен товарами, политику, культуру и т.д.;
* практическая деятельность по сбору, обработке и анализу количественных данных об общественной жизни и их публикации;
* научная дисциплина, отрасль знаний, изучающая количественную сторону массовых явлений и процессов в неразрывной связи с их количественной стороной с целью выявления закономерностей их развития. [1.2, с. 9]

Итак, математическая статистика работает там, где есть случайный эксперимент, свойства которого частично или полностью неизвестны и который мы умеем воспроизводить в одних и тех же условиях некоторое (а лучше— неограниченное) число раз.

Пусть — случайная величина, наблюдаемая в случайном эксперименте. Предполагается, что вероятностное пространство задано и не будет нас интересовать.

Проведя n раз этот эксперимент в одинаковых условиях, получим числа — значения наблюдаемой случайной величины в первом, втором и т. д. экспериментах. Случайная величина имеет некоторое распределение , которое нам частично или полностью неизвестно.

Рассмотрим подробнее набор, называемый выборкой:

.

В серии уже произведённых экспериментов выборка — это набор чисел. До того, как эксперимент проведён, имеет смысл считать выборку набором случайных величин (независимых и распределённых так же, как ). Действительно, до проведения опытов мы не можем сказать, какие значения примут элементы выборки: это будут какие-то из значений случайной величины . Поэтому имеет смысл считать, что до опыта — случайная величина, одинаково распределённая с , а после опыта — число, которое мы наблюдаем в i-м по счёту эксперименте, т. е. одно из возможных значений случайной величины . [1.3, с. 8]

Существуют два базовых типа структурированных данных: числовой и категориальный. Числовые данные имеют две формы: непрерывную и дискретную. Категориальные данные принимают только фиксированный набор значений. Двоичные данные являются важным особым случаем категориальных данных. Эти данные принимают только одно из двух значений, таких как 0/1, да/нет или истина/ложь. Еще один полезный тип категориальных данных — порядковые данные, в которых категории упорядочены; их примером является числовой рейтинг (1, 2, 3, 4 или 5). [1.4, с. 20]

Метод статистики — это целая совокупность приемов, пользуясь которыми статистика исследует свой предмет. Она включает в себя три основных метода.

Первым методом считается статистическое наблюдение, которое заключается в сборе первичного статистического материала, в научно организованной регистрации всех существующих фактов, относящихся к рассматриваемому объекту.

Второй метод называется методом группировок. Он дает возможность все собранные в результате массового статистического наблюдения все факты подвергать систематизации и классификации.

Третий метод — метод обобщающих показателей. Он позволяет характеризовать изучаемые явления и процессы при помощи статистических величин (абсолютных, относительных, средних).

Поскольку статистика имеет дело с количественными характеристиками, она широко применяет в своих исследованиях положения и методы математики. Особенно широкое применение находят в статистике теория вероятностей и математическая статистика, которые занимаются изучением абстрактных множеств единиц и действующих в них общих количественных закономерностей. Установленные этими отраслями математики законы, правила и методы статистика использует при решении своих специфических задач. В частности, важную роль играет в статистике закон больших чисел. Опираясь на закон больших чисел, статистика выявляет характерные для определенных условий закономерности, типичные количественные соотношения и уровни явлений. [1.1, с. 13]

Основными понятиями статистики являются: статистическая совокупность, единица совокупности, признак, статистический показатель, система статистических показателей.

Статистическая совокупность — это совокупность социально-экономических объектов или явлений общественной жизни, объединенных некоторой качественной основой, общей связью, но отличающимися друг от друга отдельными признаками.

Единицей совокупности называется ее первичный элемент, являющийся носителем признаков, подлежащих регистрации, и основой ведущегося при обследовании счета.

Признак статистической совокупности — качественная особенность ее единицы.

Статистический показатель — это понятие, отображающее количественные характеристики (размеры) соотношения признаков общественных явлений. Статистические показатели могут быть объемными (численность населения) и расчетными (средний возраст).

Система статистических показателей — это совокупность статистических показателей, отражающая взаимосвязи, которые объективно существуют между явлениями. Для каждой общественно-экономической формации характерна определенная система взаимосвязи общественных явлений. [1.1, с. 10]

В статистических исследованиях применяются следующие виды выборок.

Собственно-случайной, называется выборка, элементы которой отобраны без какой-либо систематизации, наудачу. Если генеральная совокупность некоторым образом упорядочена (например, она может быть упорядочена по абонентам телефонной сети), то выборка, образованная из ее элементов есть механическая. Здесь отбор осуществляется в соответствие с выбранной пропорцией через равные интервалы. В случае объединения генеральной совокупности в несколько типических групп (примерами таких групп может быть деление населения города на мужчин и женщин, или в качестве типических групп можно рассматривать граждан, не достигших трудоспособного возраста, граждан трудоспособного возраста и пенсионеров), применяют типическую выборку. Если отбор осуществляется сериями (они и являются единицами совокупности), то выборка называется серийной. Наконец, если при отборе единиц генеральной совокупности на выборку применяется несколько вышеуказанных способов (скажем, серийным и механическим способом), то выборка носит название комбинаторной. Каждый из способов отбора характеризуется своими методами и особенностями расчета числовых характеристик.

Отметим также, что отбор элементов на выборку осуществляется повторным и бесповторным способами. В первом случае, после изучения интересующих нас характеристик, элемент возвращается в генеральную совокупность и может быть вновь взят на выборку. Если отбор бесповторный, то отобранная единица генеральной совокупности в нее не возвращается и не участвует в дальнейшем отборе. [1.1, с. 24]

Статистические данные, собранные в процессе наблюдения не позволяют получить обобщающие характеристики изучаемой совокупности, выявить закономерности ее развития, так как в процессе наблюдения фиксируются характеристики только отдельных единиц совокупности.

Для получения обобщающих характеристик собранную информацию необходимо систематизировать, превратить ее в упорядоченную систему статистических показателей. Систематизация полученной информации и обобщение наблюдаемых факторов является содержанием второй стадии статистического исследования, называемой сводкой и группировкой.

Статистическая сводка представляет собой комплекс последовательных операций по обобщению конкретных единичных фактов, образующих совокупность, для выявления типичных черт и закономерностей, присущих изучаемому явлению. [1.2, с. 55]

Таким образом, основные понятия из статистики позволяют понять, как собирать, описывать и анализировать данные. Понимание этих основных понятий поможет правильно интерпретировать полученные результаты и использовать статистические методы для принятия рациональных решений в различных областях жизни. Важно помнить, что статистика не является абсолютной истиной и требует аккуратного подхода при анализе и интерпретации данных. Однако, уверенность в используемых методах и правильном выборе показателей помогут сделать более обоснованные выводы и принять соответствующие решения.

1.2 Классификация

Классификация – наиболее популярная задача машинного обучения. Она в чем-то схожа с тем, как ребенок учится определять форму и размер предметов, складывая их в раздельные кучки.

Задача классификации: предсказание категории объекта и разделение объектов согласно определенным и заданным заранее признакам. То есть машина сортирует данные по нужным категориям.

Логистическая регрессия является одним из статистических методов классификации с использованием линейного дискриминанта Фишера. Но перед ее пониманием, необходимо разобраться с базовыми регрессионными моделями.

Регрессией (уравнением регрессии) называется условное математическое ожидание Y:

,

где Y — зависимая переменная;

— независимые переменные;

— функция независимых переменных.

Таким образом, регрессия описывает поведение наблюдаемой зависимой переменной в среднем, представляя ее главную тенденцию. В связи с этим нахождение регрессии по результатам наблюдений называют сглаживанием данных.

Существуют различные регрессионные модели, определяемые выбором функции :

* простая линейная регрессия:

;

* множественная регрессия:

;

* полиномиальная регрессия:

.

Коэффициенты называются параметрами регрессии. После выбора определенной модели параметры регрессии должны быть вычислены по результатам наблюдений зависимой переменной и факторов.

В приведенные регрессионные модели параметры входят линейно. Такие модели называют линейными (по параметрам) моделями, а математические методы анализа этих моделей - линейным регрессионным анализом.

Логистическая регрессия или логит-регрессия — это статистическая модель, используемая для предсказания вероятности возникновения некоторого события путем подгонки данных к логистической кривой. [1.5]

Логистическая регрессия аналогична множественной линейной регрессии за одним исключением — исход является двоичным. При этом используются различные преобразования для того, чтобы конвертировать задачу в ту, в которой может быть подогнана линейная модель. Как и дискриминантный анализ, и в отличие от k-ближайших соседей и наивного Байеса, логистическая регрессия является подходом на основе структурированной модели нежели подходом, центрированном на данных. Указанный метод стал популярным благодаря своему высокому вычислительному быстродействию и результатам модели, которые допускают ускоренное задание баллов новым данным.

Ключевыми компонентами логистической регрессии являются функция логистического отклика и логит, в которых мы отображаем вероятность (находящуюся на шкале 0–1) в более расширенную шкалу, подходящую для линейного моделирования. Первый шаг состоит в том, чтобы думать о переменной исхода не как о двоичной метке, а как о вероятности p того, что метка равна 1. Может возникнуть соблазн наивно смоделировать p как линейную функцию предсказательных переменных:

.

Первый шаг состоит в том, чтобы думать о переменной исхода не как о двоичной метке, а как о вероятности p того, что метка равна 1. Может возникнуть соблазн наивно смоделировать p как линейную функцию предсказательных переменных:

.

Указанное преобразование обеспечивает, чтобы p оставалась между 0 и 1. Для того чтобы извлечь экспоненциальное выражение из знаменателя, мы рассматриваем перевесы вместо вероятностей. Перевесами являются отношением "успехов" (1) к "неуспехам" (0). С точки зрения вероятностей перевесы — это вероятность события, деленная на вероятность того, что событие не произойдет. Например, если вероятность, что лошадь выиграет скачки, равна 0,5, то вероятность, что "не выиграет", составит 1 - 0,5 = 0,5, и шансы равны 1,0.

*.*

Мы можем получить вероятность из шансов, используя обратную функцию перевесов:

*.*

Мы комбинируем эту формулу с функцией логистического отклика, показанной ранее, и получаем:

*.*

Наконец, взяв логарифм выражений, стоящих справа и слева от знака равенства, мы получаем выражение, которое включает линейную функцию предсказателей:

*.*

Функция логарифма шансов, также именуемая логит-функцией, отображает вероятность p из интервала (0; 1) в любое значение из интервала ().

Предсказанное значение, полученное из логистической регрессии, рассматривается с точки зрения логарифма перевесов:

*.*

Предсказанная вероятность задается функцией логистического отклика:

*.*

Линейная регрессия и логистическая регрессия имеют много схожих черт. Обе принимают параметрическую линейную форму, увязывающую предсказателей с откликом. Разведывание и отыскание лучшей модели выполняется очень похоже. Расширения линейной модели, такие как использование операции сплайнового преобразования предсказателя, одинаковым образом применимы в логистической регрессионной формулировке. Однако логистическая регрессия отличается двумя фундаментальными составляющими:

* характером выполнения подгонки модели (наименьшие квадраты не применимы);
* природой и анализом остатков от модели. [1.4, с. 228]

Регрессионный анализ позволяет дать ответ на два существенных вопроса:

* надежно ли определены параметры модели (проверка гипотезы о значимости параметров модели);
* хорошо ли описывает модель имеющиеся экспериментальные данные (проверка гипотезы об адекватности модели).

Нарушение надежности модели может проявляться и в более завуалированной форме, а именно через:

* наличие выбросов;
* наличие влиятельных наблюдений;
* нарушение основных предположений регрессионного анализа;
* наличие мультиколлинеарности.

Анализ экстремумов модели в области эксперимента дает возможность провести быструю оценку пригодности модели для предсказания. [1.6, с. 15]

В предсказательном моделировании общепринято тренировать большое число разных моделей, применять каждую к отложенной выборке и оценивать их работоспособность. Иногда после оценивания и тонкой настройки ряда моделей и, если имеется достаточно данных, третья отложенная выборка, не используемая ранее, используется для оценивания того, как выбранная модель будет работать с полностью новыми данными.

Ключевые понятия для оценивания модели:

* Точность (accuracy) — процент (или доля) случаев, классифицированных правильно;
* Матрица путаницы (confusion matrix) — Табличное изображение (2х2 в двоичном случае) количеств записей по их предсказанному и фактическому статусу классификации;
* Чувствительность (sensitivity) — Процент (или доля) всех единиц, правильно классифицированных как единицы;
* Специфичность (specificity) — Процент (или доля) всех нулей, правильно классифицированных как нули;
* ROC-кривая (ROC-curve) — График чувствительности против специфичности. [1.4, c. 223]

Можно сделать вывод, о том, что логистическая регрессия — лучший вариант для прогноза, потому что ее проще внедрять, интерпретировать и очень эффективно обучать. С ее помощью можно оценить вероятность наступления события. Она работает на принципах линейной регрессии. Именно поэтому в данной курсовой работе, для прогнозирования притока клиентов магазина, будет использоваться логистическая регрессия.

2 практическая часть

2.1 Предобработка данных

В практической части курсовой работы разработан скрипт на языке R, который с помощью модели логистической регрессии позволяет узнать вернется ли клиент в магазин.

В качестве исследуемых данных берутся данные по обращениям клиентов в магазин в период с 01.04.2020 по 30.10.2022.

Перед предобработкой данных необходимо посмотреть на их структуру. Сделать это можно с помощью встроенной в R функции «summary()». На Рисунке 1 представлен результат вызова данной функции.

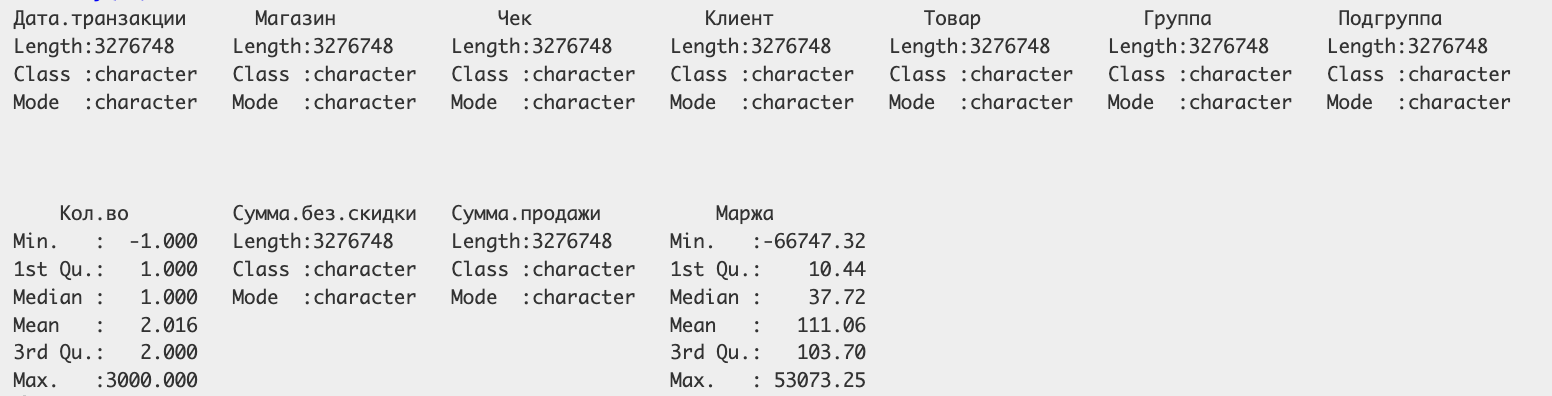


Рисунок 1 — Результат вызова функции «summary()»

Из данного рисунка видно, что объем данных довольно крупный и составляет 3276748 измерений. В исходном датафрейме 11 столбцов с данными.

Так же можно заметить, что в столбце «Кол.во», который означает кол-во товаров в чеке клиента, минимальное значение отрицательное, чего не может быть. Произведем обработку исходного датафрейма:

* сделаем столбы «Сумма.продажи», «Сумма.без.скидки» числовыми и уберем оттуда Nan значения;
* сделаем столбец «Дата.транзакции» временными;
* уберем дробные и отрицательные значения из столбца «Кол.во»;
* удалим значения в которых столбец «Сумма.без.скидки» равен нулю.

После проведения данных действий, посмотрим как выглядит наша таблица снова, с помощью функции «summary()», результат вызова которой представлен на Рисунке 2.

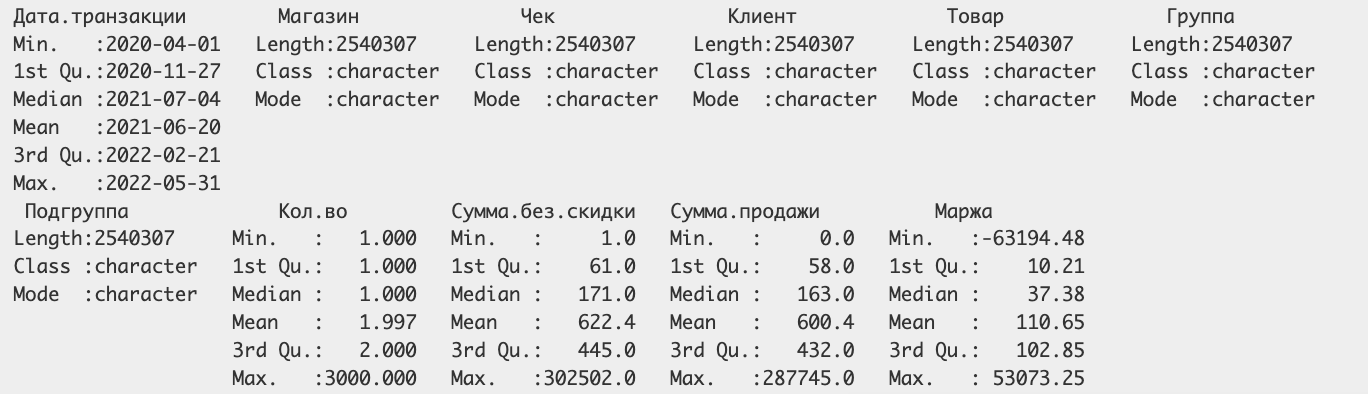


Рисунок 2 — Результат вызова функции «summary()»

Из Рисунка 2 видно, что теперь минимальное значение кол-ва товаров в чеке равно 1, а столбцы «Сумма.продажи» и «Сумма.без.скидки» стали числовыми.

Посмотрим, как выглядит первые 6 строк нашей таблицы с помощью встроенной в R функции «head()», которая на вход принимает таблицу. Результат вызова данной функции представлен на Рисунке 3.

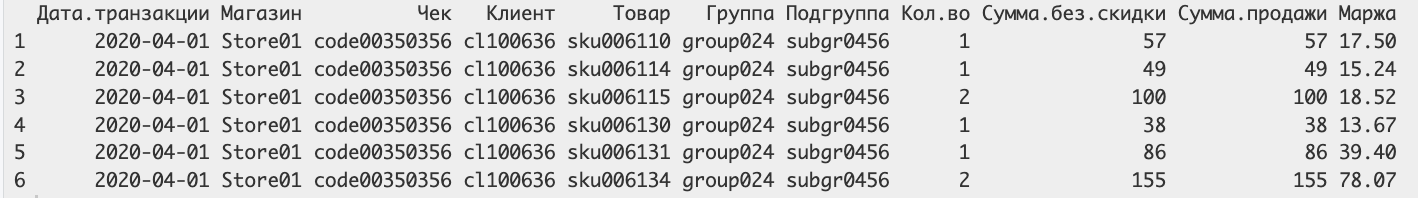


Рисунок 3 — Результат вызова функции «head ()»

Из рисунка 3 видно, что перед тем, как начать прогнозирование необходимо написать функцию «group\_by\_check» для группировки по чекам, в которой посчитаем необходимые значения:

* общую сумму продажи;
* общую маржу;
* количество позиций в чеке;
* количество уникальных позиций в чеке;
* id клиента по данному чеку;
* дату выдачи чека.

На входе функция будет принимать исходную таблицу, а на выходе возвращать отсортированную по дате выдаче чека таблицу со столбцами, описанными выше.

Затем необходимо написать функцию для группировки по клиентам. Но перед ее реализацией нужно посчитать средний интервал приходов клиента в магазин, чтобы это сделать напишем еще одну функцию «mean\_int», которая будет принимать на вход список из дат посещений клиента и возвращать средний интервал приходов.

Данная функция суммирует разницу между смежными днями и затем делит ее на общее кол-во дней.

В функции «group\_by\_client» для группировки по клиентам посчитаем необходимые значения:

* какую прибыль принес клиент;
* какой убыток принес клиент;
* кол-во его чеков;
* первое посещение;
* разницу между 1 посещением и последним;
* средний интервал между посещениями.

Данная функция возвращает таблицу со столбцами, описанными выше. После данных необходимо сделать тренировочную и тестовую выборки для будущего обучения модели логистической регрессии.

Теперь посмотрим, как выглядит наша таблица после данных группировок, с помощью встроенной в R функции «head()», которая на вход принимает нашу обработанную таблицу. Результат вызова данной функции представлен на Рисунке 4, из которого видно, что можно начать делить нашу изначальную выборку на тренировочную и тестовую.

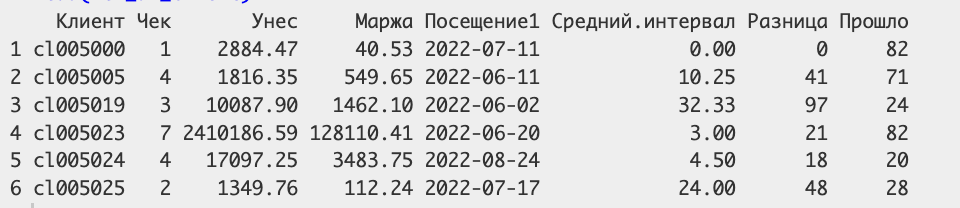


Рисунок 4 — Результат вызова функции «head ()»

Тренировочная выборка будет иметь данные за период с 01.04.2020 до 01.06.2022. Для этой выборки применим функции для группировки по чекам и клиентам. Так же необходимо узнать придет ли клиент в будущем, эти данные будут принимать значение 1 или 0, придет или не придет клиент в магазин соответственно. И возьмем мы их за период с 02.06.2020 до 01.10.2022.

Тестовая выборка будет иметь данные за период с 01.06.2022 до 01.10.2022. Для этой выборки применим функции для группировки по чекам и клиентам. И затем необходимо узнать придет ли клиент в будущем, эти данные будут принимать значение 1 или 0, придет или не придет клиент в магазин соответственно. И возьмем мы их за период с 02.06.2020 до 30.10.2022.

Таким образом, после данных действий можно приступать к созданию модели логистической регрессии и дальнейшему прогнозированию.

2.1 Прогнозирование с помощью логистической регрессии

Реализуем модель логистической регрессии в с помощью встроенной в R функции «glm()». Зависимой переменной в которой будет выступать значение 1 или 0, пришел ли клиент или нет соответственно. В качестве независимых переменных будут выступать значения:

* какую прибыль принес клиент;
* какой убыток принес клиент;
* кол-во его чеков;
* первое посещение;
* разницу между 1 посещением и последним;
* средний интервал между посещениями.

Так же необходимо одним из параметров этой функции указать «binomial», чтобы наша регрессия была логистической, и натренируем эту модель на тренировочной выборке, созданной заранее. Посмотрим на характеристику этой модели, вызвав функцию «symmary()». Данные результаты представлены на Рисунке 5.

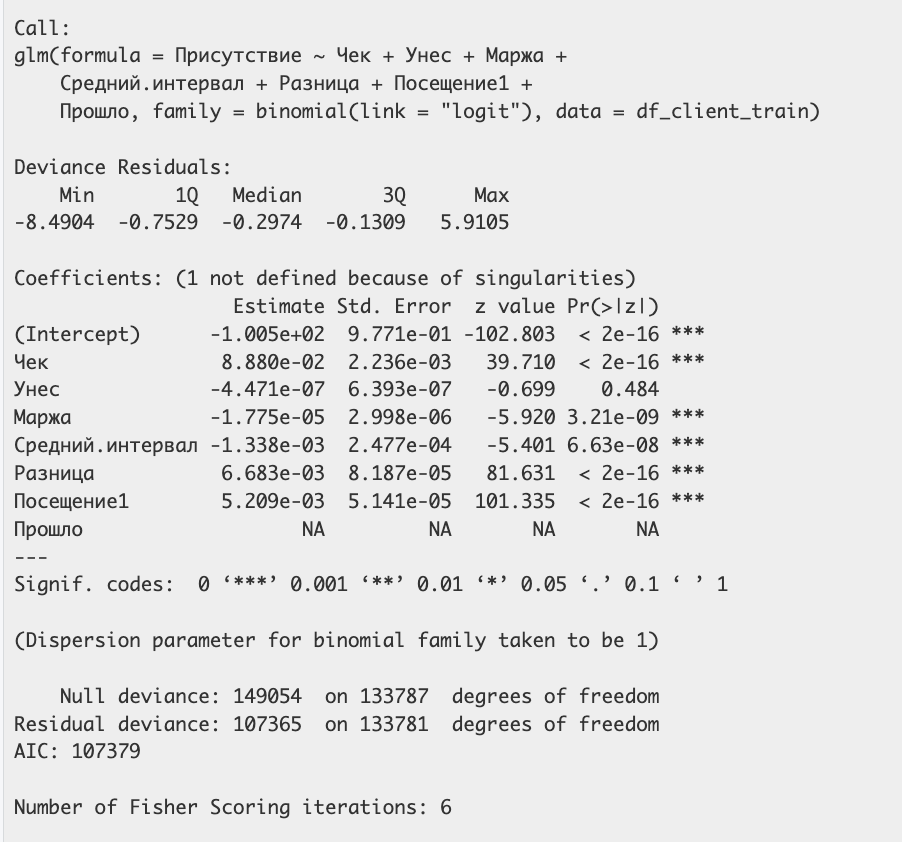


Рисунок 5 — Результат вызова функции «summary()»

Применим данную модель для тестовой выборки, созданной заранее, и сделаем прогноз с помощью встроенной в R функции «predict()», которая принимает на вход нашу модель и тестовую выборку. После прогнозирования мы получили список вероятностей, изменим его значения меньшие, чем 0.5 на 0, а большие либо равные 0,5 на 1. Выведем матрицу неточностей и другие параметры для оценки нашей модели с помощью функции «confusionMatrix()», которая принимает на вход спрогнозированные значения и значения из тестовой выборки. Результаты вызова данной функции представлены на Рисунке 6.

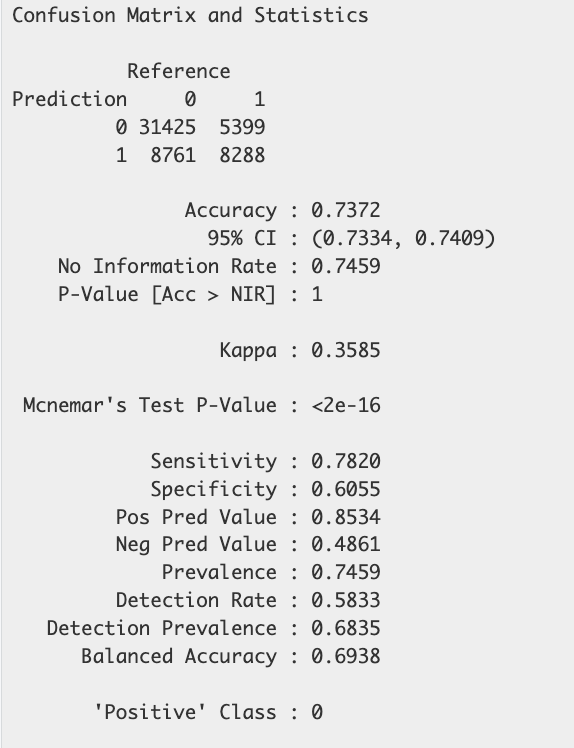


Рисунок 6 — Результат вызова функции сonfusionMatrix

Из Рисунка 6 видно, что классификатор верно отнес объект к рассматриваемому классу 31425 раз. Неверно отнес объект к рассматриваемому классу 5339 раз. Классификатор верно утверждает, что объект не принадлежит к рассматриваемому классу 8761 раз и неверно утверждает, что объект не принадлежит к рассматриваемому классу 8288 раз.

Из Рисунка 6 также видно, что точность модели составляет 0.7372, а ее 95% доверительный интервал лежит в промежутке (0.7334; 0.7409), а ее сбалансированная точность составила 0.6938. Чувствительность модели — 0.07829, а специфичность — 0.6055. Значение   
p-value вышло меньше, чем 2е-16, что значительно меньше, чем 0.05, откуда следует, что наша модель логистической регрессии статистически верна.

Далее узнаем Roc значение модели и построим график Roc кривой, представленный на Рисунке 7.

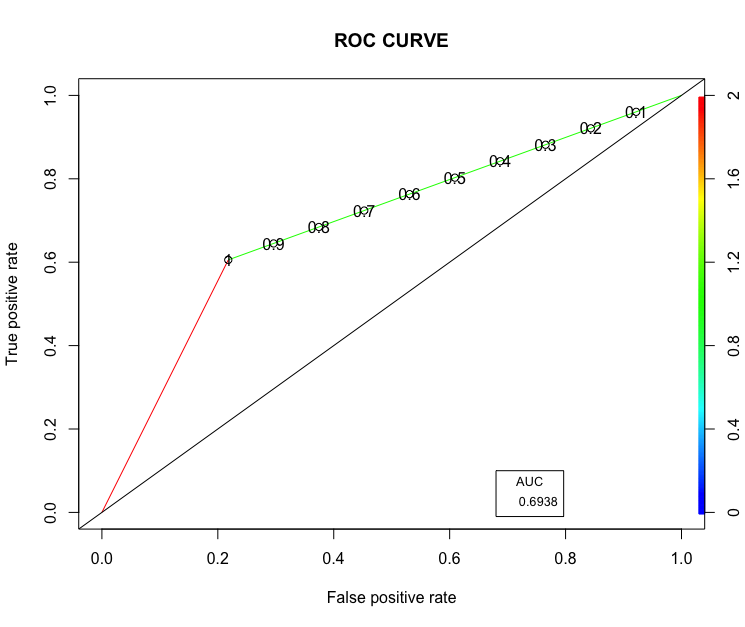


Рисунок 7 — График Roc-кривой

Из Рисунка 7 видно, что значение roc-auc составляет 0.6938, откуда следует, что качество модели относительно хорошее.

Таким образом, была проведена предобработка данных, был проведен прогноз с помощью логистической регрессии. С точностью 73% модель верно определяет рассматриваемый класс, которым выступало значение равное нулю. Ее чувствительность составила 78%, а специфичность 60%. Roc-auc значение получилось равным 0.6938, что является хорошим показателем для логистической регрессии.

Заключение

Клиентская аналитика помогает бизнесам создавать персонализированные продукты и сервисы, повышать удовлетворенность клиентов, определять Retention, улучшать маркетинг и лояльность клиентов.

Клиентская аналитика стала важной в современном бизнесе, поскольку она позволяет компаниям понимать своих клиентов и адаптироваться к потребностям рынка.

Применение клиентской аналитики можно рассматривать как обязательное условие успешного и успешно развивающегося бизнеса в нашей современной конкурентной среде.

В данной работе разработан и описан скрипт, позволяющий проанализировать и спрогнозировать возвращение клиента в магазин.

Цель данной курсовой работы — клиентский анализ, прогнозирование оттока клиентов магазина — достигнута.

В ходе курсовой работы также выполнены следующие задачи:

* изучена научная и методическая литература по математической статистике;
* использованы знания из области математической статистики с использованием современных средств обработки данных: языка программирования R;
* построена статистическая модель и ее оценка;
* прогнозирование с помощью логистической регрессии;
* изучено оформление документации.

список использованных источников

Теоретическая часть

1. Общая теория статистики: учебное пособие / Крылов В. Е.— Владимир.: Изд-во ВлГУ, 2020. — 243 с.
2. Статистика: учебное пособие / Сизова Т. М.— СПб.: Изд-во СПб НИУ ИТМО, 2013. — 176 с.
3. Математическая статистика: учебное пособие / Чернова Н. И.— Новосиб.: Изд-во гос. ун-т. Новосибирск, 2007. — 148 с.
4. Практическая статистика для специалистов Data Science /   
   Брюс. П.— СПб.: Изд-во БХВ-Петербург, 2021. — 352 с.
5. Логистическая регрессия и roc-анализ: методическое пособие / Ковалев А. А.— М.: Изд-во РГГУ, 2021. - 44 с.
6. Статистическое моделирование с использованием регрессионного анализа: электронное учебное пособие / Коновалов Ю. В.—   
   М.: Изд-во МГТУ им. Н.Э. Баумана, 2013. - 73 с
7. Статистика. Часть 1. Общая теория статистики: учеб. Пособие / Кириллов А. В.— Самара: Изд-во СГАУ, 2012. - 112 с.

ПРАКТИЧЕСКАЯ часть

1. R анализ данных и визуализация [Электронный ресурс] / Пакет caret универсальный интерфейс для доступа к десяткам алгоритмов машинного обучения. — Режим доступа: https://r-analytics.blogspot.com/2015/06/caret.html. — Дата доступа: 15.05.2023.
2. RDocumnetation [Электронный ресурс] / nnetar: Neural Network Time Series Forecasts. — Режим доступа: <https://www.rdocumentation.org/packages/forecast/versions/8.21/topics/nnetar>. — Дата доступа: 15.05.2023.
3. Бергер Е.Г. Нормоконтроль документации [Электронный ресурс]: Методические рекомендации / Бергер Е.Г., Зуев А.С. — М.: МИРЭА — Российский технологический университет, 2020. — 1 электрон. опт. диск (CD-ROM).