содержание

[введение 2](#_Toc135177157)

[1 теоретическая часть 3](#_Toc135177158)

[1.1 Основные понятие из статистики 3](#_Toc135177159)

[1.2 Регрессионный анализ 9](#_Toc135177160)

[2 практическая часть 15](#_Toc135177161)

[2.1 Предобработка данных 15](#_Toc135177162)

[2.1 Прогнозирование с помощью логистической регрессии 19](#_Toc135177163)

[заключение 22](#_Toc135177164)

[список использованных источников 23](#_Toc135177165)

[приложения 25](#_Toc135177166)

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 возвращение клиента в магазин в будущем.

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. При этом статистику прежде всего интересуют те выводы, которые можно сделать на основе анализа надлежащим образом собранных и обработанных цифровых данных.

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, как, например, скорость ветра или продолжительность времени, и дискретную, как, например, количество возникновений события. Категориальные данные принимают только фиксированный набор значений, как, например, тип экрана телевизора (плазма, LCD, LED и т. д.) или название штата (Алабама, Аляска и т. д.). 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. Закон больших чисел выражает общий принцип, в силу которого в большом числе явлений при некоторых общих условиях почти устраняется влияние случайного фактора. Достигается это в результате того, что в большом числе случаев происходит взаимопогашение индивидуальных отклонений величин одного и того же вида от общей их меры.

Опираясь на закон больших чисел, статистика выявляет характерные для определенных условий закономерности, типичные количественные соотношения и уровни явлений. [1.1, с. 13]

Главным обобщением опыта исследований любых массовых явлений служит закон больших чисел. Данный закон отражает смысл проведения исследований. Суть его в следующем. Каждое отдельное единичное явление, рассматриваемое как одно из явлений данного рода, содержит элемент случайного - оно могло быть или не быть, быть таким или иным. Однако при объединении большого числа таких явлений в общих характеристиках всей их массы случайность исчезает тем в большей мере, чем больше соединено единичных явлений, чем больше учитывается и исследуется единиц совокупности. [1.5, 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. В учебном пособии речь идет в основном о случайных и механических выборках.

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. По глубине проработки материала различают простые и сложные сводки. Простой сводкой называется операция по подсчету общих итогов по совокупности единиц наблюдения, то есть определение размера исследуемого явления. Сложной сводкой называется комплекс операций, включающих группировку единиц наблюдения, подсчет итогов по каждой группе и совокупности в целом, а также представление результатов группировки в табличной форме. [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 путѐм подгонки данных к логистической кривой. [1.7]

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 (), рис. 5.2. Цикл преобразования завершен; мы использовали линейную модель для предсказания вероятности, которую, в свою очередь, можем отобразить на метку класса путем применения правила отсечения, — любая запись с вероятностью, превышающей порог отсечения, классифицируется как 1.

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.

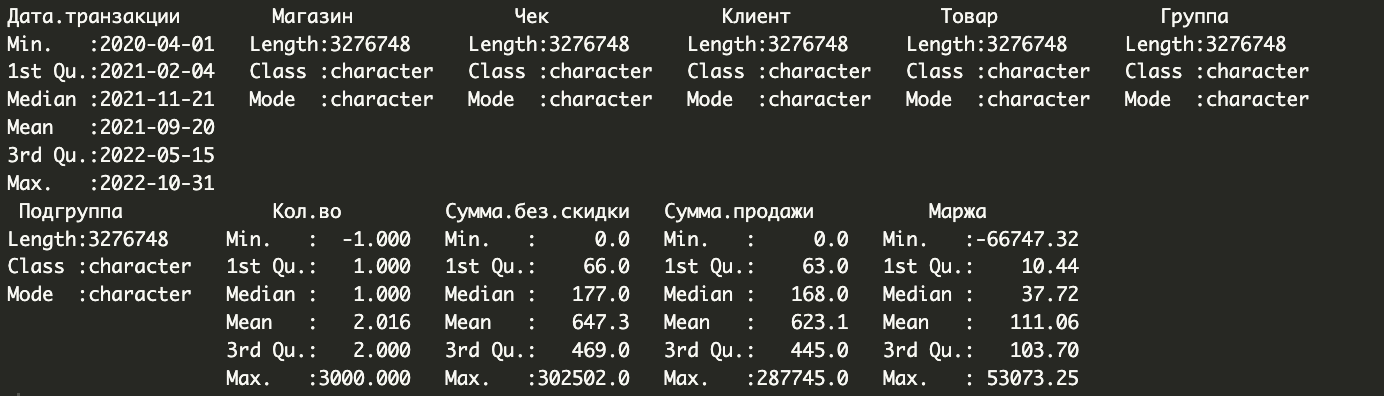


Рисунок 1 — Результат вызова функции «summary()»

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» равна нулю.

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 выходе функция возвращает 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 возвращает средний интервал приходов клиента.

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.

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.

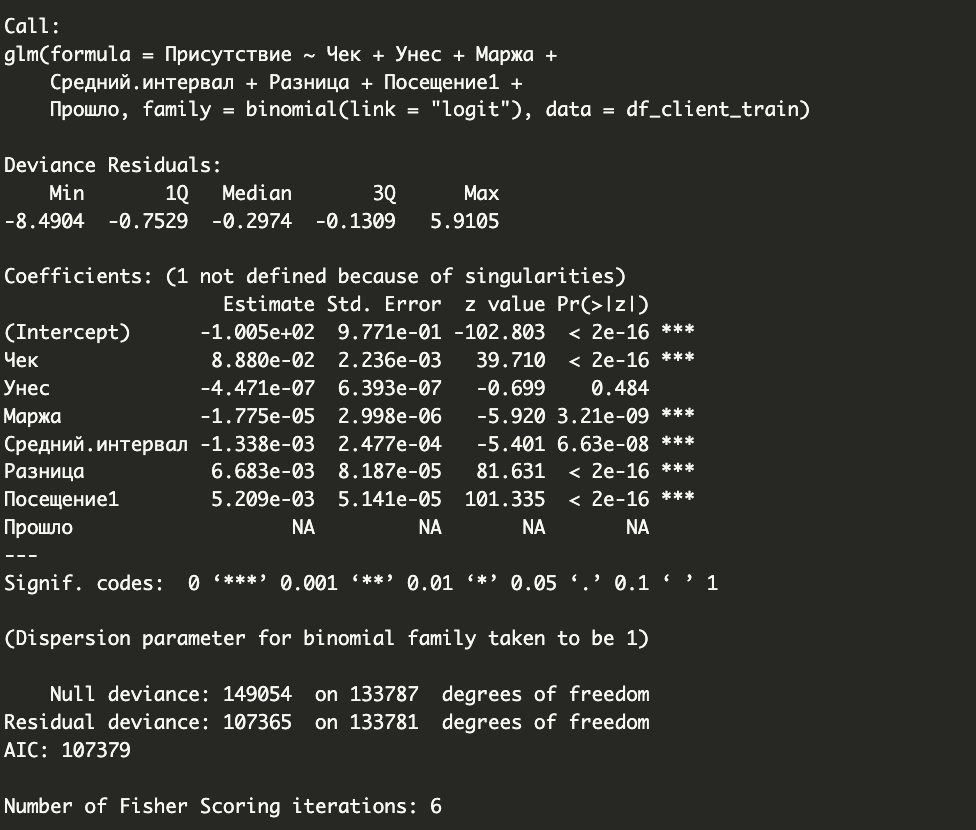


Рисунок 5 — Результат вызова функции «summary()»

Применим данную модель для тестовой выборки, созданной заранее, и оценим результаты прогноза. Выведем матрицу неточностей и другие параметры для оценки модели с помощью функции confusionMatrix, представленную на Рисунках 6-7, для нашей модели

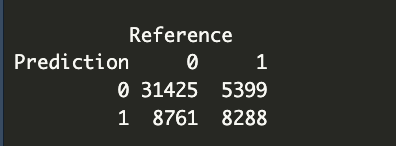


Рисунок 6 — Матрица неточностей

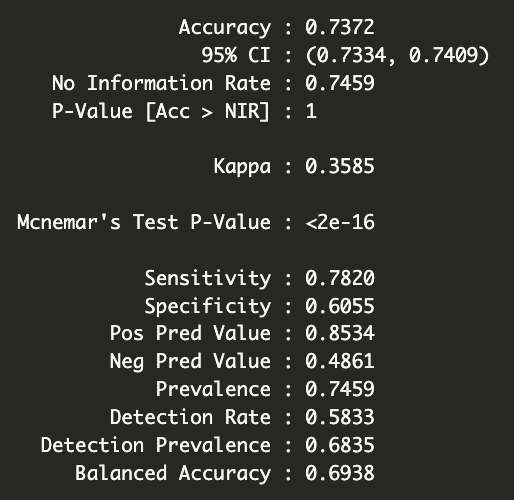


Рисунок 7 — Результат вызова функции confusionMatrix

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 7 видно, что точность модели составляет 0.7372. Чувствительность модели — 0.07829, а специфичность — 0.6055

Далее узнаем Roc значение модели и построим график Roc кривой, представленный на Рисунке 8.

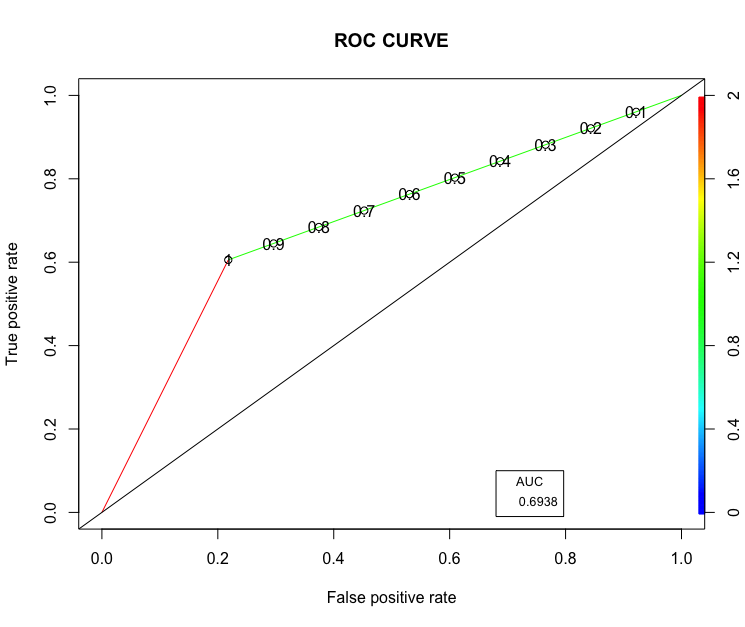


Рисунок 8— График Roc-кривой

Из Рисунка 8 видно, что значение 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% и auc значением равным 0.6938.

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 возвращение клиента в магазин в будущем — 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 их 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. Статистика. Часть 1. Общая теория статистики: учеб. Пособие / Кириллов А. В.— Самара: Изд-во СГАУ, 2012. - 112 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. Логистическая регрессия и roc-анализ: методическое пособие / Ковалев А. А.— М.: Изд-во РГГУ, 2021. - 44 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).

приложения

Приложение А — Код скрипта.

Приложение А

Листинг 1 — Код разработанного скрипта

|  |
| --- |
| date\_df<-read.csv(file='/Users/qylikys/R/Kursach/transactions.csv',  header=T, sep='\t', row.names=NULL)  # Предобработка данных ----------------------------------------------------  date\_df <- data.frame(date\_df)  head(date\_df)  df <- data.frame(date\_df)  df$Сумма.продажи <- as.numeric(gsub('NULL', '0', df$Сумма.продажи))  df$Сумма.без.скидки <- as.numeric(gsub('NULL', '0', df$Сумма.без.скидки))  df$Дата.транзакции <- as.Date(df$Дата.транзакции, format="%d/%m/%Y")  prelast\_df <- df  last\_df <- df  n <- nrow(df)  if(!("dplyr" %in% installed.packages())){  install.packages("dplyr")  }  library(dplyr)  prelast\_month <- as.Date('01/06/2022', format = "%d/%m/%Y")  last\_month <- as.Date('01/10/2022', format = "%d/%m/%Y")  df <- filter(df,  Дата.транзакции < prelast\_month &  Кол.во - as.integer(Кол.во) == 0 &  Сумма.без.скидки != 0)  # Функции для группировок -------------------------------------------------  group\_by\_check <- function(df){ #Функция группировки по чеку  sale\_sum <- aggregate(df, Сумма.продажи ~ Чек, FUN=sum) #Общая сумма продажи  marja\_sum <- aggregate(df, Маржа ~ Чек, FUN=sum) #Общая маржа  count\_pos <- aggregate(df, Кол.во ~ Чек, FUN=sum) #Кол-во позиций в чеке  unique\_pos <- summarise(group\_by(df, Чек),  Уникальные.товары=length(unique(Товар))) #Уникальные  client <- summarise(group\_by(df, Чек), Клиент=Клиент[1]) #Клиенt  date <- summarise(group\_by(df, Чек), Дата.транзакции=Дата.транзакции[1])  df\_check <- data.frame(date[2],  sale\_sum[1],  client[2],  unique\_pos[2],  count\_pos[2],  sale\_sum[2],  marja\_sum[2])  df\_check <- df\_check[order(df\_check$Дата.транзакции,  decreasing = F), ]  return(df\_check)  }  #Функция для рассчета среднего интервала ммжду месяцами  mean\_int <- function(arr){  summ = 0  n = length(arr)  if (n != 1){  for (i in (1:(n-1))){  day\_diff = as.numeric(gsub(' days', '',  as.Date(arr[i+1],  format = "%d/%m/%Y") -  as.Date(arr[i],  format = "%d/%m/%Y")))  summ = summ + day\_diff  }  }  else{  summ = 0  }  return(signif(summ / n, digits = 4))  }  #Функция для группировки по клиентам  group\_by\_client <- function(df\_check, for\_last\_month){  sale\_sum <- aggregate(df\_check, Маржа ~ Клиент, FUN=sum) #Принес  loss\_sum <- summarise(group\_by(df\_check, Клиент), Унес=  (sum(Сумма.продажи) - sum(Маржа))) #Унес  count\_check <- aggregate(df\_check, Чек ~ Клиент, FUN=length) #Кол-во чеков  first\_date <- summarise(group\_by(df\_check, Клиент), Посещение1=  Дата.транзакции[1]) #Первое посещение  diff\_date <- summarise(group\_by(df\_check, Клиент), Разница= as.numeric(  gsub(' days', '',  Дата.транзакции[length(Дата.транзакции)] -  Дата.транзакции[1]))) #Разница посл. и 1 дня  passed\_time <- summarise(group\_by(df\_check, Клиент), Прошло= as.numeric(  gsub(' days', '',  for\_last\_month -  Дата.транзакции[length(Дата.транзакции)]))) #Прошло дней  date\_int <- summarise(group\_by(df\_check, Клиент), Средний.интервал=  mean\_int(Дата.транзакции))  df\_client <- data.frame(sale\_sum[1],  count\_check[2],  loss\_sum[2],  sale\_sum[2],  first\_date[2],  date\_int[2],  diff\_date[2],  passed\_time[2]  )  return(df\_client)  }  # Группировки -------------------------------------------------------------  #Группировки  df\_check <- group\_by\_check(df) #Группировка по чекам  df\_client <- group\_by\_client(df\_check, prelast\_month) #Группировка по клиентам  # Создание модели ---------------------------------------------------------  date\_df$Дата.транзакции <- as.Date(date\_df$Дата.транзакции, format="%d/%m/%Y")  date\_df$Сумма.продажи <- as.numeric(gsub('NULL', '0', date\_df$Сумма.продажи))  date\_df$Сумма.без.скидки <- as.numeric(gsub('NULL', '0',  date\_df$Сумма.без.скидки))  df\_train <- filter(date\_df,  Дата.транзакции >=  prelast\_month &  Дата.транзакции <  last\_month &  Кол.во - as.integer(Кол.во) == 0 &  Сумма.без.скидки != 0)  new\_month <- data.frame(Клиент = unique(df\_train$Клиент), Присутствие = 1)  df\_client\_train <- dplyr::left\_join(df\_client, new\_month, by = "Клиент")  df\_client\_train[is.na(df\_client\_train)] <- 0  model <- glm(Присутствие ~  Чек +  Унес +  Маржа +  Средний.интервал +  Разница +  Посещение1 +  Прошло,  data=df\_client\_train, family = binomial(link="logit"))  summary(model)  # Группировки для основного периода  new\_df <- data.frame(prelast\_df)  new\_df$Сумма.продажи <- as.numeric(gsub('NULL', '0', new\_df$Сумма.продажи))  new\_df$Сумма.без.скидки <- as.numeric(gsub('NULL', '0', new\_df$Сумма.без.скидки))  new\_df$Дата.транзакции <- as.Date(new\_df$Дата.транзакции, format="%d/%m/%Y")  new\_df <- filter(new\_df,  Дата.транзакции > prelast\_month &  Дата.транзакции < last\_month &  Кол.во - as.integer(Кол.во) == 0 &  Сумма.без.скидки != 0)  new\_df\_check <- group\_by\_check(new\_df)  new\_df\_client <- group\_by\_client(new\_df\_check, last\_month)  # Группируем клиентов  future\_month <- as.Date('01/11/2022', format = "%d/%m/%Y")  last\_df <- data.frame(last\_df)  last\_df$Дата.транзакции <- as.Date(last\_df$Дата.транзакции, format="%d/%m/%Y")  last\_df <- filter(last\_df,  Дата.транзакции >= last\_month &  Кол.во - as.integer(Кол.во) == 0 &  Сумма.без.скидки != 0)  temp\_df <- summarise(group\_by(last\_df, Клиент), Присутствие=1)  # Прогнозируем на последний месяц -----------------------------------------  new\_df\_client\_test <- dplyr::left\_join(new\_df\_client, temp\_df, by = "Клиент")  new\_df\_client\_test[is.na(new\_df\_client\_test)] <- 0  predictResult <- predict(model, newdata = new\_df\_client\_test, type="response")  predictResult <- ifelse(predictResult >= 0.5, 1, 0 )  #Сравниваем результат прогнозирования  #Матрица неточностей  library(caret)  confusionMatrix(factor(predictResult), factor(new\_df\_client\_test$Присутствие))  #Точность модели  missing\_classerr <- mean(predictResult != new\_df\_client\_test$Присутствие)  print(paste( 'Accuracy =' , 1 - missing\_classerr))  if(!("caTools" %in% installed.packages())){  install.packages("caTools")  }  library(caTools)  if(!("ROCR" %in% installed.packages())){  install.packages("ROCR")  }  library(ROCR)  # Кривая ROC-AUC  ROCPred <- prediction(predictResult, new\_df\_client\_test$Присутствие)  ROCPer <- performance(ROCPred, measure = "tpr" ,  x.measure = "fpr" )  auc <- performance(ROCPred, measure = "auc" )  auc <-auc@y.values[[1]]  auc  # Построение кривой  plot(ROCPer)  plot(ROCPer, colorize = TRUE,  print.cutoffs.at = seq( 0.1 , by = 0.1 ),  main = "ROC CURVE" )  abline(a = 0 , b = 1 )  auc <- round(auc, 4)  legend(.68 , .10 , auc, title = "AUC" , cex = 0.8 ) |