**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 5](#_Toc132897341)

[1 Теоретическая часть 6](#_Toc132897342)

[1.1 Факторный анализ 6](#_Toc132897343)

[1.2 Детерминированный анализ 13](#_Toc132897344)

[2 Практическая часть 17](#_Toc132897345)

[2.1 Метод главных компонент 17](#_Toc132897346)

[2.2 Проведение факторного анализа 21](#_Toc132897347)

[Заключение 26](#_Toc132897348)

[Список использованных источников 27](#_Toc132897349)

[Приложения 30](#_Toc132897350)

# ВВЕДЕНИЕ

Одна из проблем в сфере услуг – оттек клиентов, многие клиенты покупают абонементы, но перестают посещать заведения, и отменяют их до истечения срока. Для того чтобы понять, что заставляет пользователей уйти, и выбрать стратегию по работе с клиентами, можно провести анализ данных клиентов и выявить факторы, которые наибольшим образом влияют на отказ клиента от абонемента.

Цель работы – провести анализ данных клиентов и провести факторный анализ для определения факторов, которые наибольшим образом влияют на отток клиентов.

Задачи, решаемые в данной курсовой работе:

* изучение научной и методической литературы по исследуемой проблеме;
* использование знаний математической статистики с использованием метода главных компонент для проведения факторного анализа;
* построение модели машинного обучения для возможности прогнозирования;
* обучение качественному оформлению документации.

Данная курсовая работа является актуальной в многих областях, включая бизнес, маркетинг и науку о здоровье. Факторный анализ может помочь фитнес-центрам разработать более эффективные предложения на основе потребностей клиентов, а также оптимизировать свои услуги и повысить качество обслуживания своих клиентов.

# 1 Теоретическая часть

## 1.1 Факторный анализ

Факторный анализ – это метод многомерного статистического анализа данных, который позволяет определить наименьшее число факторов, необходимых для объяснения наибольшей части вариации исходных переменных. Факторы – это ошибки или переменные, которые раскрывают скрытые связи между наблюдаемыми переменными. Этот метод помогает нам понять, какие переменные взаимосвязаны между собой и как они связаны с общими факторами [1.1].

Основными целями факторного анализа являются:

* идентификация скрытых факторов: факторный анализ помогает увидеть скрытые факторы, которые не учитываются в оригинальных данных;
* сокращение размерности: факторный анализ позволяет свести большое количество переменных к небольшому количеству факторов;
* упрощение множественных данных: факторный анализ помогает свести множество переменных к нескольким базовым факторам, что упрощает процесс анализа данных и понимания сути проблемы;
* поиск ассоциаций: факторный анализ может помочь определить связи между разными переменными и выявить структуру данных.

Существуют следующие типы факторного анализа:

* + - 1. Детерминированный (функциональный) – результативный показатель представлен в виде произведения, частного или алгебраической суммы факторов.
      2. Стохастический (корреляционный) – связь между результативным и факторными показателями является неполной или вероятностной.
      3. Прямой (дедуктивный) – от общего к частному.
      4. Обратный (индуктивный) – от частного к общему.
      5. Одноступенчатый и многоступенчатый.
      6. Статический и динамический.
      7. Ретроспективный и перспективный.

Также факторный анализ может быть разведочным – он осуществляется при исследовании скрытой факторной структуры без предположения о числе факторов и их нагрузках и конфирматорным, предназначенным для проверки гипотез о числе факторов и их нагрузках. Практическое выполнение факторного анализа начинается с проверки его условий.

Обязательные условия факторного анализа:

* все признаки должны быть количественными;
* число признаков должно быть в два раза больше числа переменных;
* выборка должна быть однородна;
* исходные переменные должны быть распределены симметрично;
* факторный анализ осуществляется по коррелирующим переменным.

При анализе в один фактор объединяются сильно коррелирующие между собой переменные, как следствие происходит перераспределение дисперсии между компонентами и получается максимально простая и наглядная структура факторов. После объединения коррелированность компонент внутри каждого фактора между собой будет выше, чем их коррелированность с компонентами из других факторов.

Выделим основные этапы факторного анализа:

* + - 1. Подготовка данных. На этом этапе происходит сбор данных, их обработка и проверка на достоверность. Также необходимо провести предварительный анализ данных, чтобы понять, какие переменные будут участвовать в факторном анализе.
      2. Выбор метода факторного анализа. В зависимости от типа данных и цели исследования можно выбрать различные методы факторного анализа, например, главные компоненты, метод максимального правдоподобия или метод наименьших квадратов.
      3. Определение количества факторов. С помощью факторного анализа необходимо выделить наиболее значимые факторы и определить их количество. Для этого можно использовать различные критерии, например, критерий Кайзера или скрин-тест.
      4. Расчет факторной матрицы. На этом этапе происходит расчет матрицы факторных нагрузок, которая показывает связь между переменными и факторами. Уже на этом этапе можно сделать выводы о том, какие переменные наиболее важны для каждого фактора.
      5. Интерпретация результатов. По завершении факторного анализа необходимо проанализировать полученные результаты и сделать выводы о том, какие факторы наиболее значимы и какие переменные вносят наибольший вклад в каждый фактор. Это позволяет лучше понимать структуру данных и принимать более обоснованные решения.

Факторный анализ может применяться как для количественных, так и для качественных данных. Однако, есть некоторые отличия и особенности в использовании этого метода для разных типов данных. Количественный факторный анализ является более распространенным и используется для исследования множества переменных, измеренных с помощью числовых данных. В этом случае, данные должны быть непрерывными и могут быть представлены в виде чисел, таблиц или графиков. Анализ основывается на расчете корреляционной матрицы между переменными и на поиске латентных факторов, объясняющих наибольшую долю дисперсии в исходных данных.

Качественный факторный анализ используется, когда данные представлены в виде категорий или номинальных значений. В этом случае, каждая переменная может быть представлена в виде качественной, категориальной переменной. Анализ базируется на расчете матрицы частотности для определения взаимосвязей между переменными и на поиске факторов, объясняющих вариативность в качественных данных. Особенность качественного факторного анализа заключается в использовании методов анализа категорий для обнаружения скрытых факторов.

По характеру взаимосвязи между показателями различают [методы детерминированного](https://afdanalyse.ru/publ/finansovyj_analiz/1/determinirovannyj_faktornyj_analiz/11-1-0-101) и [стохастического факторного анализа](https://afdanalyse.ru/publ/finansovyj_analiz/1/stokhasticheskij_faktornyj_analiz/11-1-0-103).

Детерминированный факторный анализ представляет собой методику исследования влияния факторов, связь которых с результативным показателем носит функциональный характер, т. е. когда результативный показатель факторной модели представлен в виде произведения, частного или алгебраической суммы факторов.

Методы детерминированного факторного анализа:

* метод цепных подстановок;
* метод абсолютных разниц;
* метод относительных разниц;
* интегральный метод;
* метод логарифмирования.

Данный вид факторного анализа наиболее распространен, поскольку, будучи достаточно простым в применении (по сравнению со стохастическим анализом), позволяет осознать логику действия основных факторов развития предприятия, количественно оценить их влияние, понять, какие факторы, и в какой пропорции возможно и целесообразно изменить для повышения эффективности производства.

Стохастический анализ представляет собой методику исследования факторов, связь которых с результативным показателем в отличие от функциональной является неполной, вероятностной (корреляционной). Если при функциональной (полной) зависимости с изменением аргумента всегда происходит соответствующее изменение функции, то при корреляционной связи изменение аргумента может дать несколько значений прироста функции в зависимости от сочетания других факторов, определяющих данный показатель.

Методы стохастического факторного анализа используются для изучения взаимосвязи между множеством переменных, называемых факторами. Они позволяют определить общие факторы, которые могут влиять на эти переменные и выявляют зависимости между ними. Каждый из методов стохастического факторного анализа использует свои математические формулы.

Среди методов стохастического факторного анализа можно выделить следующие:

* + - 1. Способ парной корреляции – основан на оценке связи между переменными. Данный метод позволяет выявить наличие корреляции между двумя факторами.

,

где – коэффициент корреляции,

– переменные.

* + - 1. Множественный корреляционный анализ – используется для оценки связи между одной зависимой переменной и двумя или более независимыми переменными. Этот метод позволяет определить, какие факторы влияют на зависимую переменную с наибольшей силой.

,

где – зависимая переменная,

– независимые переменные,

– свободный член,

– коэффициенты регрессии,

– случайная ошибка.

* + - 1. Матричные модели – представляют собой компьютерную обработку данных, основанную на матричных операциях. Они используются для определения общих факторов, а также для выявления нескольких существующих групп переменных.

Они описываются следующей формулой:

,

где – матрица наблюдений (переменных),

– матрица факторных коэффициентов,

– матрица факторов,

– матрица ошибок.

* + - 1. Математическое программирование – метод, который основывается на строгом математическом анализе факторов. Он позволяет выявлять такие факторы, которые наиболее эффективно действуют на рассматриваемые переменные.

Она записывается следующей формулой:

при условиях:

,

где – коэффициенты целевой функции,

– переменные,

– коэффициенты ограничений на переменные ,

– значения, до которых должны ограничиваться значения левых частей уравнений в ограничениях.

* + - 1. Метод исследования операций – это метод, используемый для определения наилучшего решения в сложных системах, которые содержат множество переменных. Для исследования операций используются различные модели, например, сетевые модели. Одна из таких моделей – сетевой график.

Он записывается следующей формулой:

,

где – наибольшее время, нужное для выполнения работы ,

– продолжительность работы от до .

* + - 1. Теория игр – набор математических моделей, которые используются для изучения различных оптимальных стратегий в условиях неопределенности. Одной из наиболее распространенных формул является формула Нэша.

Она записывается следующим образом:

,

где – оптимальная стратегия игрока 1,

– вероятность выбора -ой стратегии игроком 1,

– выигрыш игрока 1 при -ой и -ой стратегии игроков 1 и 2 соответственно.

Необходимо также различать статический и динамический факторный анализ. Первый вид применяется при изучении влияния факторов на результативные показатели на соответствующую дату. Другой вид представляет собой методику исследования причинно-следственных связей в динамике.

## 1.2 Детерминированный анализ

Детерминированный факторный анализ (Deterministic Factor Analysis) – это метод анализа данных, направленный на выявление латентных (скрытых) факторов, объясняющих совокупную вариабельность наблюдаемых переменных. В отличие от классического факторного анализа, при котором факторы определяются исключительно статистически, в детерминированном факторном анализе рассматриваются случаи, когда количество факторов известно заранее.

Алгоритм детерминированного факторного анализа основан на линейной регрессии, когда каждая наблюдаемая переменная является линейной комбинацией проверяемых факторов. Таким образом, этот метод позволяет не только выявить латентные факторы, но и оценить вклад каждого из них в объяснение наблюдаемой переменной.

При этом стандартные предположения о нормальности распределений и постоянстве условной дисперсии нарушаются, что требует использования методов оценки параметров, аналогичных методу наименьших квадратов с помощью максимального правдоподобия.

Для проведения детерминированного факторного анализа должны быть выполнены следующие условия:

* + - 1. Линейность и нормальность: переменные должны быть линейными и нормальными. Это означает, что отклонения от среднего значения должны иметь нормальное распределение.
      2. Крупномасштабность: количество переменных должно быть значительным по сравнению с количеством наблюдений. Обычно используют от 5 до 30 переменных, но это зависит от конкретной задачи.
      3. Независимость: переменные должны быть независимы друг от друга. Это означает, что одна переменная не должна быть линейной комбинацией другой.
      4. Отсутствие мультиколлинеарности: переменные не должны быть сильно связаны друг с другом, чтобы избежать проблемы мультиколлинеарности, которая может привести к неустойчивости оценок параметров и неинтерпретируемым результатам.
      5. Квадратичная согласованность: ковариационная матрица переменных должна быть квадратично-согласованной, что означает, что она должна быть симметричной, неотрицательно определенной и единичной диагональю.
      6. Правильный выбор метода оценки факторов: для детерминированного факторного анализа обычно используют метод главных компонент или метод максимального правдоподобия.

Если все эти условия выполнены, то детерминированный факторный анализ может быть использован для определения наиболее значимых факторов, которые объясняют наблюдаемую вариативность в данных.

Выделим основные шаги детерминированного факторного анализа:

* + - 1. Предварительная обработка данных, включающая стандартизацию (центрирование и масштабирование) переменных.
      2. Определение параметров модели, включая количество факторов, которые должны быть заданы заранее, и функцию связи между латентными факторами и наблюдаемыми переменными, которая может быть линейной или нелинейной.
      3. Оценка параметров модели с использованием метода максимального правдоподобия.
      4. Оценка факторных нагрузок и коэффициентов регрессии наблюдаемых переменных на латентные факторы для интерпретации результатов.

Существует несколько методов детерминированного факторного анализа, каждый из которых может дать различные результаты:

* + - 1. Метод главных компонент (Principal Component Analysis) – это метод, который позволяет определить наименьшее количество факторов, которые могут объяснить максимальную долю общей вариации. Факторы, которые образуются, являются линейными комбинациями исходных переменных.
      2. Метод наименьших квадратов (Least Squares method) – это метод, который также позволяет определить наименьшее количество факторов, но в отличие от метода главных компонент, он подбирает такие факторы, которые лучше всего соответствуют наблюдаемым данным.
      3. Обобщенный метод наименьших квадратов (Generalized Least Squares method) – это расширение метода наименьших квадратов, которое учитывает корреляции между переменными и может использоваться в случае, когда данные имеют неодинаковую дисперсию.
      4. Метод максимального правдоподобия (Maximum Likelihood method) – это метод, который основан на предположении о распределении данных и позволяет оценить параметры распределения для каждого фактора.

Каждый из методов факторного анализа может быть использован в зависимости от специфики исследования и имеет свои преимущества и недостатки. При выборе метода необходимо учитывать характеристики данных и задачу, которую необходимо решить.

Одной из особенностей факторного анализа является то, что он основывается на предположении о наличии латентных переменных, которые не наблюдаются напрямую, но влияют на исходные данные. Поэтому, для правильного применения факторного анализа необходимо убедиться в корректности выбора переменных и определении правильного количества факторов. В зависимости от количества учитываемых факторов принято выделять коэффициенты парной корреляции (однофакторная зависимость) и множественной корреляции (двух- и более факторная зависимость).

# 2 Практическая часть

## 2.1 Метод главных компонент

В курсовой работе будут рассмотрены данные фитнес-центра. Одна из проблем в сфере услуг – оттек клиентов, многие клиенты покупают абонементы, но перестают посещать заведения, и отменяют их до истечения срока. Для борьбы с оттоком клиентов можно провести анализ данные клиентов и выявить факторы, которые наибольшим образом влияют на отказ клиента от абонемента.

Нам известны данные клиентов одного из фитнес-центров, которые имеют следующие признаки (Рисунок 1):

* + - 1. Row – индекс строки данных.
      2. Registration – уникальный номер клиента.
      3. Zipcode – индекс клиента.
      4. Age – возраст.
      5. Partner\_company – продан ли абонемент по партнерской компании.
      6. Friend\_promo – продан ли абонемент по рекомендации друга.
      7. Contract\_period – продолжительность подписки в месяцах.
      8. Lifetime – продолжительность истории посещений.
      9. Class\_registation\_weekly – среднее число занятий в неделю.
      10. Avg\_additional\_charges\_total – средняя плата за дополнительные услуги.
      11. Cancellation\_freq – среднее число отмен и пропусков занятий.
      12. Exited – аннулирован ли абонемент.

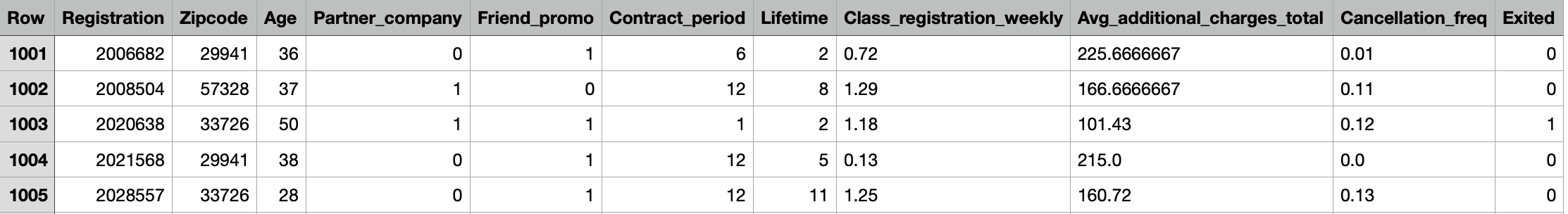


Рисунок – Данные фитнес-центра

Определение необходимого количества факторов в факторном анализе весьма важно, поскольку это определяет, как много факторов нужно выделить из данных. Найденное количество факторов также может оказать влияние на итоговые результаты анализа.

Существует несколько различных методов для определения необходимого количества факторов для факторного анализа. Используем два метода и сравним полученные количества факторов: метод критерия Кайзера и метод экранного теста.

Метод критерия Кайзера основан на показателе, называемом главных компонент. В этом методе необходимо выбрать все главные компоненты с собственными значениями больше 1. Таким образом, количество факторов может быть определено как количество главных компонент с собственными значениями, большими 1.

Метод главных компонент может быть использован для сокращения размерности данных и упрощения сложной информации в наборах данных. Принцип работы метода главных компонент основан на нахождении линейных комбинаций исходных переменных, которые максимально сохраняют исходную информацию. Данные факторы называются главными компонентами. Первая главная компонента соответствует наивысшей дисперсии из исходных переменных, вторая – следующей по величине дисперсии и так далее.

Преимущество компонентного анализа заключается в том, что он не требует знания теоретических связей между переменными, а также предоставляет возможность снижения шума, сглаживания и улучшения качества модели. Однако, при использовании метода главных компонент необходимо помнить о том, что он может исказить некоторые аспекты исходных данных, а также что полученные факторы являются линейными комбинациями начальных переменных, что ограничивает их интерпретацию. Также метод главных компонент чувствителен к выбросам, что может привести к искажению результатов.

Метод главных компонент включает в себя следующие шаги:

1. Центрирование данных. На первом этапе проводится центрирование данных. Это означает, что из каждого значения переменной вычитается среднее значение этой переменной по всей выборке. Такой подход гарантирует, что каждая переменная имеет среднее значение равное 0.
2. Рассчет ковариационной матрицы. Рассчитывается ковариационная матрица, которая характеризует степень связи между переменными. В ковариационной матрице элементы на главной диагонали показывают вариацию каждой переменной, а остальные элементы показывают степень корреляции между переменными.
3. Вычисление собственных значений и собственных векторов. Собственные значения и собственные векторы используются для определения главных компонент. Собственные значения показывают, какую долю объясненной дисперсии описывает каждая главная компонента. Собственные векторы могут быть использованы для вычисления новых ортогональных измерений – главных компонент.
4. Преобразование данных. Исходные данные преобразовываются путем умножения каждой строки (то есть каждого наблюдения) на матрицу собственных векторов, чтобы получить новые переменные – главные компоненты.
5. Выбор числа главных компонент. Выбирается наименьшее количество главных компонент, которые объясняют заданную долю общего объясненного разброса. В выборе количества компонент могут помочь различные критерии, такие как "локоть", "ломаную" или "график собственных значений".
6. Интерпретация главных компонент. Последний этап заключается в интерпретации главных компонент. Это может быть сложно, так как каждый фактор представляет собой линейную комбинацию многих исходных переменных. Но на основе значений коэффициентов наиболее выдающихся переменных вектора факторной нагрузки можно сделать выводы об основных тенденциях.

Проведем его на языке программирования R. Для начала рассчитаем главные компоненты с собственными значениями больше 1. Теперь количество факторов может быть определено как количество главных компонент с собственными значениями, большими 1. Получим результат, показанный на рисунке 2.



Рисунок – Количество факторов методов Кайзера

Рассмотрим еще один метод определения количества факторов – метод экранного теста. Метод экранного теста также использует главные компоненты, но в качестве критерия используется график "экран" или скрининг-график, на котором отображаются собственные значения главных компонент. Количество факторов выбирается на основе "критического значения", которое может быть определено путем анализа склонности собственных значений на графике экрана.

Рассчитаем собственные значения и выведем график на экран (Рисунок 3). На данном графике по оси OY изображены собственные значения главных факторов и по оси OX номера факторов.

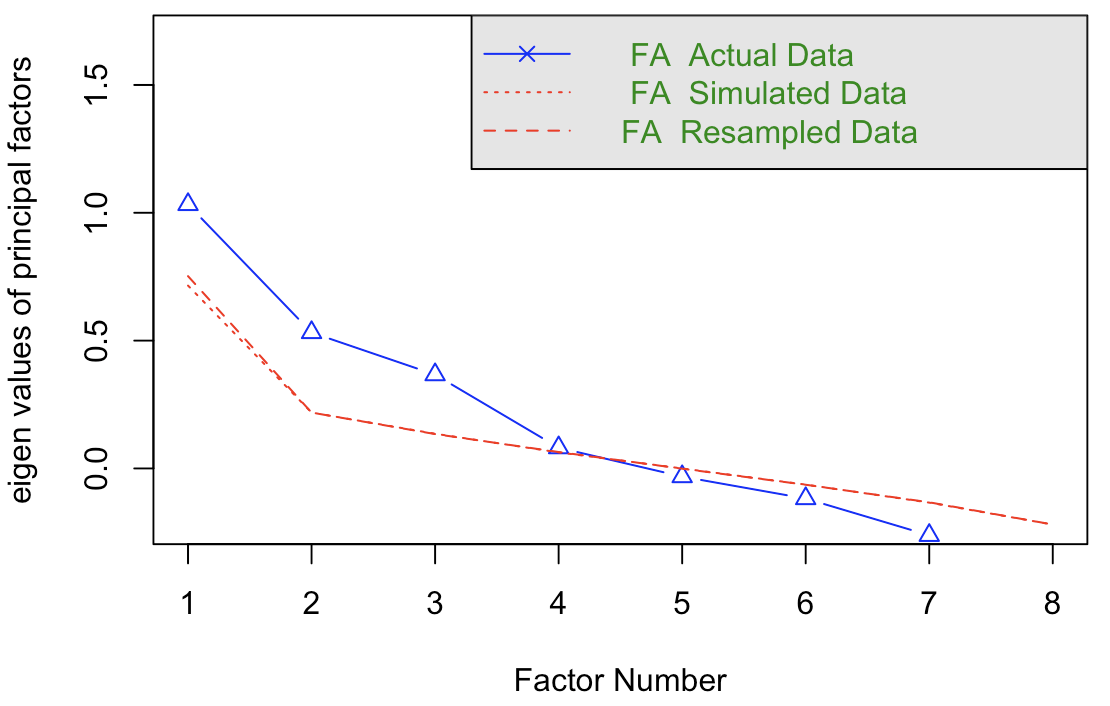


Рисунок – Метод экранного теста

На графике показаны как фактически и смоделированные данные, что позволяет нам сделать вывод о количестве необходимых факторов для проведение факторного анализа. Из данного делаем вывод о том, что необходимое число факторов равно 4, так как про увеличении числа факторов точность модели не будет критически изменяться.

Таким образом оба метода для определения количества факторов для данных фитнес-центра показали число факторов равное 4, это число факторов используем на практике для проведения факторного анализа.

## 2.2 Проведение факторного анализа

Применим к данным факторный анализ с ограничением до 4 факторов с помощью функции factanal(), в которую передаем данные и ограничения. Выведем результаты анализа. Первое, что выводит данная функция это параметр Uniquenesses, этот параметр показывает, какую долю общей вариации данных объясняет каждый выбранный фактор (Рисунок 4). Как правило, выбираются факторы, которые объясняют большую часть вариации, но не обязательно все факторы объясняют большую вариацию.

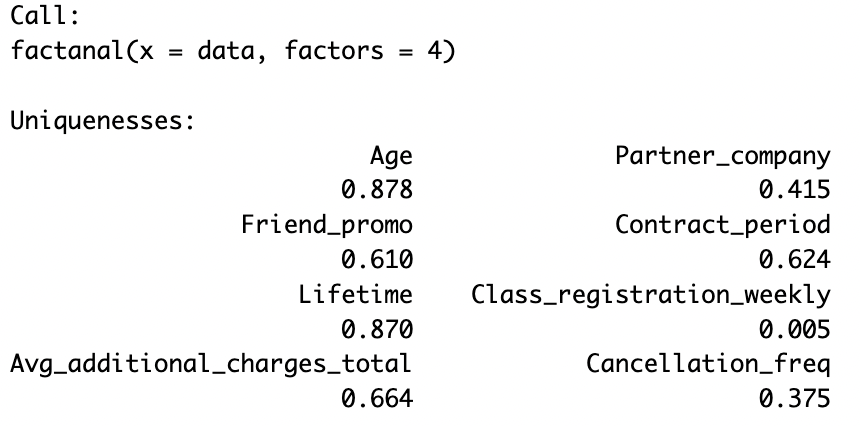


Рисунок – Вывод параметра Uniquenesses

Далее функция выводит корреляционную матрицу (Рисунок 5). Корреляционная матрица: это матрица, в которой вычисляются коэффициенты корреляции между исходными переменными и факторами. Она показывает, как каждая переменная связана с каждым фактором.

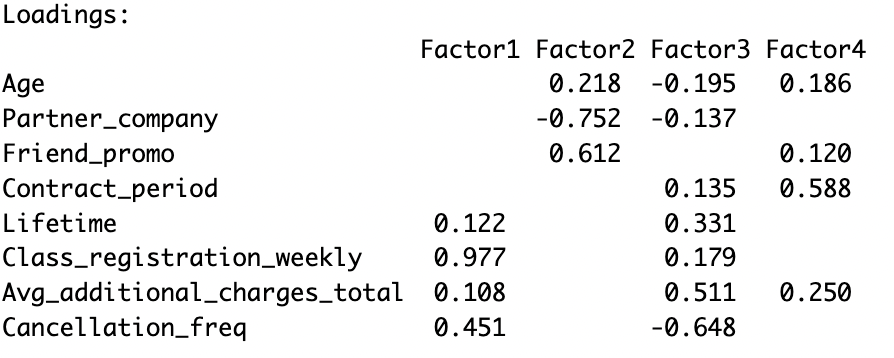


Рисунок – Корреляционная матрица

Чтобы проанализировать эту матрицу, можно визуализировать ее при помощи корреляционных графиков (Рисунок 6). На данном графике по оси OX расположены факторы, а по оси OY параметры, влияющие на факторы. Цвет показывает собственные значения, которые имеют значения в промежутке от -1 до 1. Чем больше величина собственного значения, тем большее влияние он оказывает на фактор.

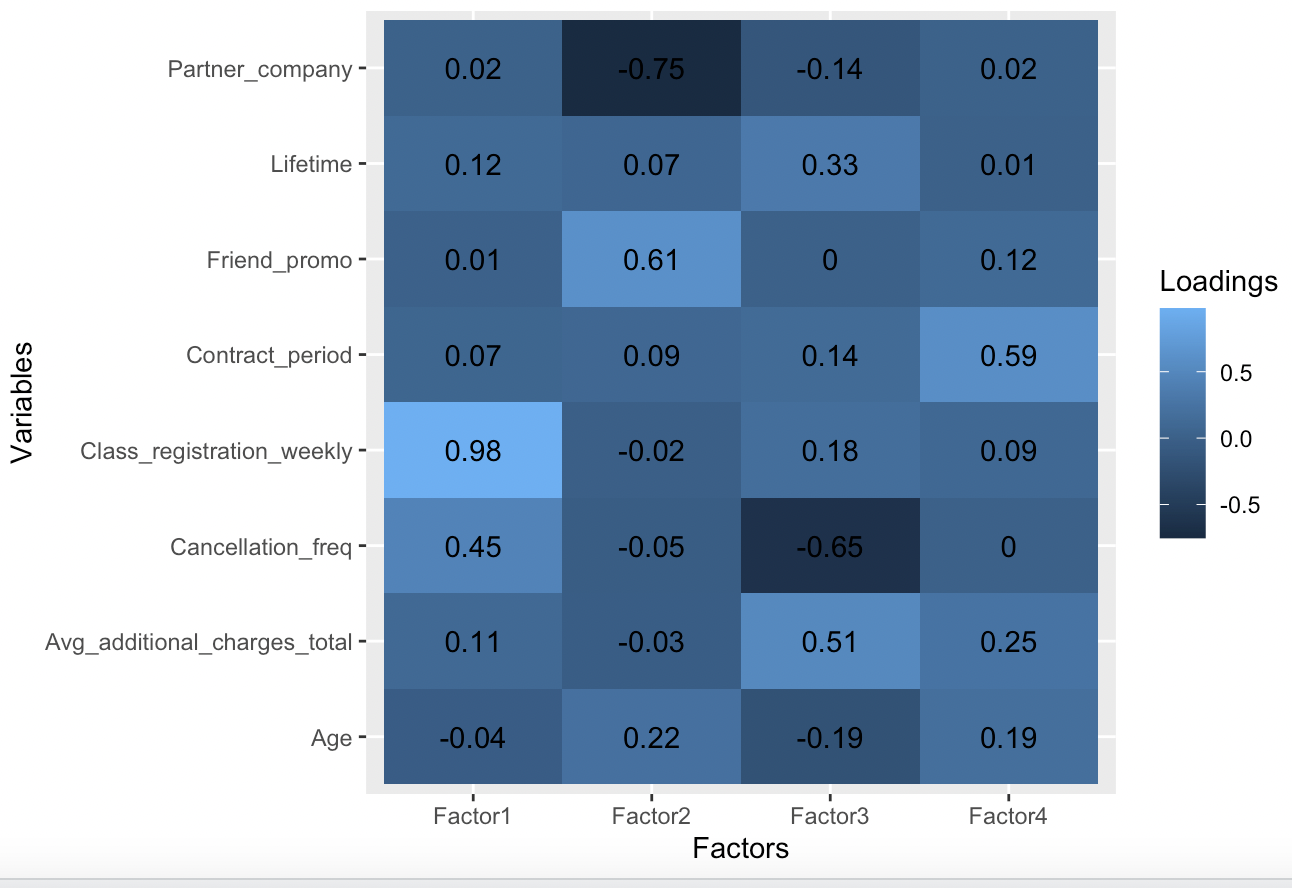


Рисунок – Визуализация корреляционной матрицы

Как видно из данной матрицы на первый фактор наибольшим образом влияет параметр Class\_registation\_weekly (среднее число занятий в неделю), на второй фактор – Friend\_promo (продан ли абонемент по рекомендации друга), на третий – Avg\_additional\_charges\_total (средняя плата за дополнительные услуги), на четвертый – Contract\_period (продолжительность подписки в месяцах).

Также выводятся общие результаты работы функции (Рисунок 7).

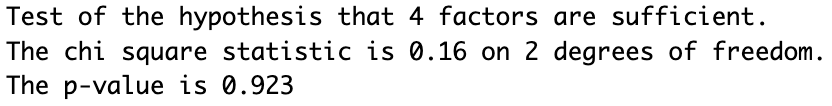


Рисунок – Результат факторного анализа

Данный результат проведения факторного анализа указывает на то, что существует значительная вероятность того, что 4 фактора являются достаточными, чтобы описать исходные данные. В данном случае, проведенный тест гипотезы состоит в проверке того, что число факторов, выбранных для моделирования данных, является оптимальным.

Значение Chi square statistic – это статистическая характеристика, которая показывает, как хорошо выбранные факторы соответствуют имеющимся данным, где меньшее значение указывает на лучшее соответствие. Значение 0.16 на 2 градусах свободы (degrees of freedom) говорит о том, что выбранные факторы хорошо соответствуют данным.

Значение p-value – это вероятность получить такое или еще более значимое различие между ожидаемыми и наблюдаемыми значениями, если нулевая гипотеза верна. В данном случае, p-value равно 0.923, что говорит о том, что нулевую гипотезу не удалось отвергнуть. Это указывает на то, что выбранные 4 факторы достаточны для моделирования данных.

Таким образом, результаты теста подтверждают, что выбранные 4 фактора являются оптимальными и достаточными для описания данных.

Теперь построим Байесовский классификатор. Байесовский классификатор – это алгоритм машинного обучения, который основывается на теореме Байеса для определения вероятности принадлежности объекта к определенному классу. Он используется для решения задачи классификации, когда необходимо определить к какому из предопределенных классов относится данный объект, исходя из некоторых характеристик этого объекта.

Для работы байесовского классификатора, сначала необходимо построить модель, основанную на обучающей выборке данных. Эта выборка должна включать объекты, уже отнесенные к определенным классам, и информацию об их характеристиках (признаках). Данные характеристики используются для определения априорных вероятностей каждого класса.

Затем, при поступлении нового объекта, байесовский классификатор анализирует его характеристики и вычисляет апостериорные вероятности принадлежности объекта к каждому из классов. Для этого он использует формулу Байеса, которая позволяет пересчитать априорную вероятность класса при условии новой информации. Итоговый класс объекта определяется как класс с наибольшей апостериорной вероятностью. Для уменьшения шума и повышения точности классификации, часто используется метод наивного Байеса, предполагающий, что все признаки объекта независимы друг от друга.

После построения и получения предсказаний модели оценим модель с помощью функции confusionMatrix. Функция confusionMatrix в R является частью библиотеки caret, которая используется для оценки качества работы модели машинного обучения. Данная функция позволяет построить матрицу ошибок (confusion matrix), а также рассчитать метрики качества модели, такие как точность, полнота, F-мера и другие.

Данная функция выводит confusion matrix – матрицу ошибок, которая позволяет оценить, насколько точно модель классификации предсказывает классы. В данном случае, модель предсказывает два класса, обозначенные как 0 и 1 (Рисунок 8).

Для класса 0 модель предсказала правильно 18 раз, но 2 раза ошиблась (Predicted 1), а для класса 1 модель предсказала правильно 1 раз, но 0 раз ошиблась.

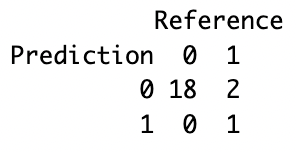


Рисунок – Матрица ошибок

Результат работы функции показывает, что модель справилась с задачей на 90.48% точности (Accuracy), а значение Kappa равно 0.4615, что является низким показателем (Рисунок 9).

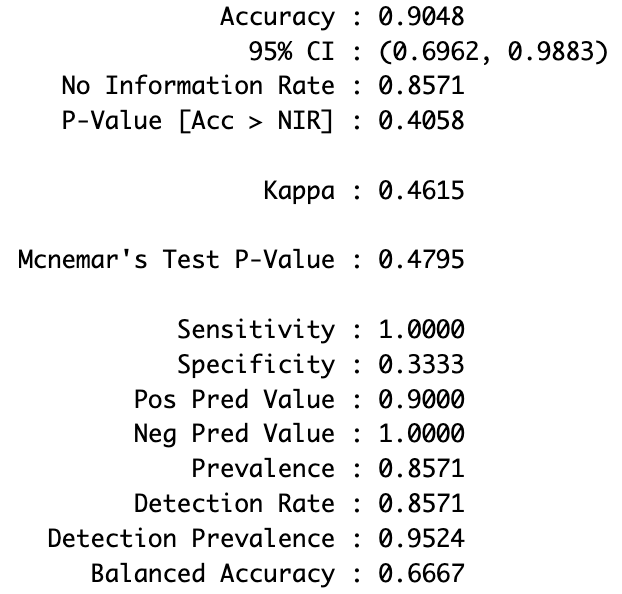


Рисунок – Результат моделирования

Значение Specificity равно 0.3333, что говорит о том, что модель плохо справляется с определением класса 1, т.к. правильно определяет его только в 33.33% случаев (хотя это и единственный раз, когда она делает это правильно). Neg Pred Value - значение отрицательного предиктивного значения равно 1, то есть если модель предсказала класс 1, то она всегда ошибалась. Pos Pred Value - значение положительного предиктивного значения для класса 0 равно 0.9, что говорит о том, что если модель предсказала класс 0, то в 90% случаев она права.

# Заключение

Одна из проблем в сфере услуг – оттек клиентов, многие клиенты покупают абонементы, но перестают посещать заведения, и отменяют их до истечения срока. Для того чтобы понять, что заставляет пользователей уйти, и выбрать стратегию по работе с клиентами, можно провести анализ данных клиентов и выявить факторы, которые наибольшим образом влияют на отказ клиента от абонемента.

В результате работы проведен анализ данных клиентов и факторный анализ для определения факторов, которые наибольшим образом влияют на отток клиентов. Выявлены основные факторы, с помощью которых была построена модель машинного обучения для прогнозирования.

В ходе работы выполнены следующие задачи:

* изучена научная и методическая литературы по исследуемой проблеме;
* использованы знания математической статистики и метод главных компонент для проведения факторного анализа;
* проведен факторный анализ;
* построена модели машинного обучения;
* пройдено обучение качественному оформлению документации.

# Список использованных источников

Теоретическая часть

1. Макшанов А. В. «Технологии интеллектуального анализа данных»: учебное пособие / А. В. Макшанов, А. Е. Журавлев. — 2-е изд., стер. Санкт-Петербург: Лань, 2022. — 212 с.: ил. — (Учебники для вузов. Специальная литература). — Москва: Финансы и Статистика, 2002. — 480 с.
2. Шовин, В.А. «ФАКТОРНЫЙ АНАЛИЗ КАЧЕСТВЕННЫХ ПОКАЗАТЕЛЕЙ» / В. А. Шовин // Математические структуры и моделирование. — 2019. — № 3. — С. 88-96. — Текст: электронный // Лань: электронно-библиотечная система.
3. Терешина Н.П., «Основы факторного анализа грузооборота на железнодорожном транспорте»: учебное пособие / Н.П. Терешина, В. А. Подсорин, М. Г. Данилина [и др.]. – Москва: РУТ (МИИТ), 2019 202 c. Текст: электронный // Лань: электронно-библиотечная система.
4. Габдуллин, Н. М. «Развитие человеческого капитала и цифровой экономики в регионах России: факторный и кластерный анализ»: монография / Н. М. Габдуллин. - Казань: КФУ, 2019. — 268 с. - Текст: электронный // Лань: электронно- библиотечная система.
5. Прохорова, И. С. «Анализ хозяйственной деятельности в инновационной сфере»: учебное пособие / И. С. Прохорова. - Москва: РУТ (МИИТ), 2019. - 201 с. – Текст: электронный // Лань: электронно- библиотечная система
6. Бариленко В. И. «Новые тенденции в развитии информационно- аналитического обеспечения и контрольных процессов в современной экономике: сборник научных трудов» / под редакцией В. И. Бариленко. - Москва: Финансовый университет, 2015. - 267 с. Текст: электронный // Лань: - электронно-библиотечная система.
7. Ганичева А.В. «Прикладная статистика»: учебное пособие / А.В. Ганичева. — Санкт-Петербург: Лань, 2022. — с. 88.
8. Федорова Н.П., Миронова З.А. «Статистика. Общая теория статистики»: учебное пособие / Н.П. Федорова, З.А. Миронова. — Ижевск: Ижевская ГСХА, 2019. — с. 37.
9. Карабутов, Н.Н. «Введение в теорию эксперимента в исследовании систем» / Н. Н. Карабутов. - Санкт-Петербург: Лань, 2023. — 168 с. Текст: электронный // Лань: электронно-библиотечная система.
10. Круценюк К.Ю. «Корреляционно-регрессионный анализ в эконометрических моделях»: учебное пособие / К.Ю. Круценюк. — Норильск: НГИИ, 2018. — с. 6.
11. Митина Т.В. «Многомерные случайные величины. Корреляционный анализ»: учебное пособие / Т.В. Митина. — Дубна: Государственный университет «Дубна», 2021. — с. 27.

Практическая часть

1. Ганичева А.В. «Прикладная статистика»: учебное пособие / А.В. Ганичева. — Санкт-Петербург: Лань, 2022. — с. 88.
2. Карабутов, Н.Н. «Введение в теорию эксперимента в исследовании систем» / Н. Н. Карабутов. - Санкт-Петербург: Лань, 2023. — 168 с. Текст: электронный // Лань: электронно-библиотечная система.
3. Ахмадиев Ф.Г. «Математическое моделирование и методы оптимизации»: Учебное пособие / Ф.Г. Ахмадиев, Р.М. Гильфанов. Казань: Изд-во Казанск. гос. архитект.-строит. ун-та, 2017. - 178 с.
4. RPubs [Электронный ресурс] / Визуализация описательных статистик. – Режим доступа: <https://rpubs.com/AllaT/sem17-04-2017>
5. Математика и статистика [Электронный ресурс] / Проверка гипотез: концепция p-value. – Режим доступа: <http://math-hse.info/f/2017-18/ps-ms/hypo-test-pvalue.pdf>
6. Geeksforgeeks [Электронный ресурс] / Logistic Regression in R Programming. – Режим доступа: <https://www.geeksforgeeks.org/logistic-regression-in-r-programming/amp/>
7. Cran.r [Электронный ресурс] / psych: Procedures for Psychological, Psychometric, and Personality Research. – Режим доступа: https://cran.r-project.org/web/packages/psych/index.html
8. Rstudio-pubs-static [Электронный ресурс] / Анализ данных в R. – Режим доступа: https://rstudio-pubs-static.s3.amazonaws.com/820737\_d2991cdf8a774863a707cdc1ce922f67.html

# Приложения

Приложение А — Программный код.

**Приложение А**

install.packages('corrplot')

install.packages('psych')

install.packages('ggplot2')

install.packages("gmodels")

install.packages('caret')

install.packages('e1071')

library(corrplot)

library(psych)

library(ggplot2)

library(gmodels)

library(dplyr)

#library('caret')

library('e1071')

df = read.csv(file='/Users/anastasiyarutkovskaya/Downloads/gym.csv', header = TRUE, sep=",", dec='.', fill=TRUE, comment.char="")

head(df[,4:11]) # выведем первые 6 строк

data<-df[,4:11]

str(data) # выведем информацию о типах данных и переменных

summary(data) # выведем статистические данные о каждой переменной

# расчет главных компонент

fit <- princomp(data)

# Выводим результаты анализа

summary(fit)

# расчет главных компонент с собственными значениями более 1

NumOfFactors <- sum(fit$sdev > 1)

# вывод результата

cat("Количество факторов, определенное методом Кайзера:", NumOfFactors)

# расчет собственных значений

fact <- fa.parallel(data, n.iter=500, fm="ml", fa="fa")

# вывод результата

cat("Количество факторов, определенное методом экранного теста:", fact$parallel[1])

# Применяем факторный анализ с ограничением до 4 факторов

fa\_result <- factanal(data, factors = 4)

# Выводим результаты анализа

print(fa\_result)

#Получим вектор относительных весов

loadings<-fa\_result$loadings

# Определяем данные

dframe <- data.frame(var = rownames(loadings), factor = rep(colnames(loadings), each = nrow(loadings)), value = as.vector(loadings))

# Строим тепловую карту

ggplot(dframe, aes(x = factor, y = var, fill = value)) +

geom\_tile() +

geom\_text(aes(label=round(value, 2))) +

labs(x = "Factors", y = "Variables", fill = "Loadings")

# Выводим полученные факторы и их вклад в общую дисперсию

fa\_result$loadings

# Добавляем значения факторов в исходный датасет

factor1 <- df$Class\_registration\_weekly

factor2 <- df$Friend\_promo

factor3 <- df$Avg\_additional\_charges\_total

factor4 <- df$Contract\_period

newdata = data.frame(factor1,factor2,factor3,factor4,df$Exited)

newdata

#Разделим данные на обучающую и тестовую выборки

index = sample(2,nrow(newdata),prob = c(0.9,0.1),replace=TRUE)

set.seed(1234)

train = newdata[index==1,]

test = newdata[index==2,]

test\_data = test[1:4]

test\_data

test\_label = test[,5]

test\_label

# Обучим модель:

vector <-newdata[1:dim(train),]

vector

train$factor1

vector$df.Exited

model=naiveBayes(vector$df.Exited~.,train)

model

# Получим предсказания модели:

test\_result=predict(model,test\_data)

test\_data

test\_result

# Оценка модели:

caret::confusionMatrix(factor(test\_result),factor(test\_label))