Федеральное государственное образовательное бюджетное

учреждение высшего образования

**«Финансовый университет при Правительстве Российской Федерации»**

**(Финансовый университет)**

Факультет информационных технологий и анализа больших данных

Департамент анализа данных и машинного обучения

Курсовая работа по дисциплине: «Технология анализа данных и машинного обучения»

на тему:

«**Машинное обучение в задачах предсказания оттока клиентов**»

Вид исследуемых данных:

Данные клиентов телекоммуникационной компании

Направление подготовки 01.03.02 «Прикладная математика и информатика»

Профиль «Анализ данных и принятие решений в экономике и финансах»

Выполнила:  
студентка группы

ПМ19-3

Филимонова Ю. М.

Руководитель:  
доцент  
Болтачев Э.Ф.

Москва 2022

**Оглавление**

[**ВВЕДЕНИЕ** 3](#_Toc121939218)

[**Глава 1. Теоретическая часть** 4](#_Toc121939219)

[**1.1. Использование машинного обучения в задачах прогнозирования оттока клиентов** 4](#_Toc121939220)

[**1.1.1 Отток клиентов и его влияние на бизнес** 4](#_Toc121939221)

[**1.1.2 Прогнозирование оттока и его важность** 6](#_Toc121939222)

[**1.1.3 Причины использования машинного обучения для прогнозирования оттока** 7](#_Toc121939223)

[**1.2. Модели машинного обучения для решения задачи классификации** 8](#_Toc121939224)

[**1.2.1. Логистическая регрессия** 9](#_Toc121939225)

[**1.2.2. Метод опорных векторов** 11](#_Toc121939226)

[**1.2.3. Алгоритм k-ближайших соседей (KNN)** 13](#_Toc121939227)

[**1.2.4 Дерево решений** 14](#_Toc121939228)

[**1.2.5. Случайный лес** 16](#_Toc121939229)

[**1.2.6. Наивный байесовский алгоритм** 17](#_Toc121939230)

[**1.2.7. Градиентный бустинг** 18](#_Toc121939231)

[**1.3. Оценка моделей классификации** 19](#_Toc121939232)

[**Глава 2. Практическая часть** 22](#_Toc121939233)

[**2.1. Выбор набора данных для анализа** 22](#_Toc121939234)

[**2.2. Предварительный анализ данных** 23](#_Toc121939235)

[**2.3. Построение моделей** 31](#_Toc121939236)

[**ЗАКЛЮЧЕНИЕ** 36](#_Toc121939237)

[**СПИСОК ЛИТЕРАТУРЫ** 38](#_Toc121939238)

[**ПРИЛОЖЕНИЕ** 40](#_Toc121939239)

# **ВВЕДЕНИЕ**

Для бизнеса каждый клиент очень важен. На рынке постоянно появляются новые конкуренты, желающие предложить людям инновационные и выгодные услуги, способные переманить чужих клиентов к себе. Если смотреть на эту ситуацию со стороны потребителя, то такой расклад, очевидно, благоприятный: можно выбирать удобные для себя условия, получать дополнительные опции или хорошие скидки на услуги. Если же что-то потребителю не понравится, то множество других компаний будут рады новому клиенту. Но смотря на такую ситуацию со стороны продавца, стоит понимать, что Ваш бизнес должен предполагать длительные отношения с клиентами, потеря которых способна нанести значительный ущерб. А поскольку привлечение новых клиентов дороже и сложнее, чем удержание нынешних, то для наилучшего удержание существующих клиентов, компаниям необходимо своевременно определять группу лиц, желающих перестать пользоваться их услугами.

Компании, которые имеют огромную клиентскую базу и работают с большими данными, пользуются возможностями машинного обучения и интеллектуального анализа данных. От поведения клиентов зависит дальнейшая судьба компании. Именно поэтому не сложно догадаться, что в сфере бизнеса наиболее полезным примером использования машинного обучения является прогнозирование оттока клиентов.

В данный момент существует множество работ и исследований, прогнозирующие поведение клиентов, поскольку эта проблема является важной в области CRM, что в свою очередь подтверждает актуальность работы.

Целью работы является определение наилучших моделей машинного обучения для быстрого и эффективного прогнозирования оттока клиентов в компании.

К основным задачам, необходимым для достижения цели, относятся:

1. изучение понятия оттока клиентов и определение влияние оттока на судьбу компании;

2. обзор основных методов машинного обучения для решения задачи классификации

3. рассмотрение нескольких алгоритмов машинного обучения, которые могут использоваться для прогнозирования оттока клиентов;

4. создание наилучшей модели прогнозирования оттока, использую набор реальных данных.

Объектом исследования является данные людей, пользующихся услугами телекоммуникационной компании, учитывая некоторые сведения о них: их демографические характеристики, услуги, которыми они пользуются, длительности пользования услугами оператора, методы оплаты и т.д.

Предметом исследования выступает анализ алгоритмов машинного обучения для создания эффективной модели прогнозирования оттока клиентов в компании.

# **Глава 1. Теоретическая часть**

# **1.1. Использование машинного обучения в задачах прогнозирования оттока клиентов**

## **1.1.1 Отток клиентов и его влияние на бизнес**

Отток клиентов подразумевает предрасположенность клиентов к прекращению использования услуг кампании и к прекращению это быть платежеспособным клиентом какого-либо бизнеса. Тот процент людей, которые перестали использовать продукт, предлагаемый кампанией, в течение определенного периода времени, называется показателем оттока клиентов.

У оттока клиентов может быть много потенциальных причин, в том числе:

* Плохое обслуживание клиентов
* Несоответствие рыночному качеству и стандартам
* Недостаточно значения
* Плохая клиентоориентированность
* Клиенты нашли другие альтернативные решения
* Несовместимость с другими системами
* Высокая цена
* Отсутствие инноваций
* Плохой пользовательский интерфейс/UX
* Ненадежность

Отток негативно влияет на каждую область бизнеса, включая оценку, ежемесячный доход, стоимость привлечения клиентов и темпы роста.

Когда клиенты уходят, компания не только теряет деньги в краткосрочной перспективе, но и теряет долю рынка в пользу конкурентов, особенно если эти клиенты являются постоянными покупателями, которые имеют большую лояльность. Это означает, что у компании как у бизнеса меньше доля рынка и меньше возможностей для роста. Очевидным результатом этого является резкое падение оценки компании.

Конечно, когда происходит потеря существующего клиента, теряется будущий доход. Это, в свою очередь, отражается на долгосрочной жизнеспособности и прибыльности бизнеса. Клиент, которого потеряли сегодня, возможно, был постоянным покупателем, лояльным к бренду, и, таким образом, представляет собой значительную потерю.

Кроме того, они могут быть заменены новым клиентом, у которого ранее не было отношений с компанией, а это означает, что приобретение этого нового клиента будет стоить больше с точки зрения ресурсов и затрат на маркетинг, чем удержание существующего клиента. Другими словами, стоимость привлечения клиентов увеличивается.

Наконец, отток снижает масштабируемость и рост. Если определенное количество клиентов покидает компанию, перестает пользоваться продуктом или услугой каждый месяц, компания не сможете масштабироваться с той же скоростью, что и его конкуренты. Также можно понять, что потребуется больше времени для достижения критической массы и получения прибыли, что ограничит потенциал роста.

## **1.1.2 Прогнозирование оттока и его важность**

Предсказание оттока — это метод предиктивной аналитики, который предсказывает, когда клиенты могут покинуть вашу компанию. Это важный инструмент для бизнеса по нескольким причинам:

* Это помогает выявить потенциальные риски
* Это позволяет предприятиям принимать превентивные меры
* Это помогает лучше понять клиентов, тем самым облегчая поддержание выгодных отношений с клиентами.
* Это помогает принимать лучшие бизнес-решения и повышать рентабельность инвестиций.

Рассмотрим подробно каждую из этих причин.

Прогнозирование оттока выявляет риски до того, как они произойдут. Если бизнес знает, когда он потеряет клиентов, он может принять меры до того, как клиент откажется от продукта или услуги. Это особенно верно для стартапов и малых предприятий, где каждый доллар на счету — время, потраченное на выявление рисков на раннем этапе, может быть потрачено с пользой для предотвращения полной реализации этих рисков.

Прогнозируя отток, компания может планировать заранее и предотвращать ненужные расходы, а также предотвращать проблемы, которые могут возникнуть позже, если не будут предприняты действия. Например, если бизнес знает, что он потеряет постоянных клиентов, потому что у них есть другие варианты, чем с компанией, он может инвестировать в предоставление альтернативных решений до того, как клиенты уйдут.

В качестве альтернативы, если бизнес определил, что его клиенты перестанут использовать определенные функции его продукта или услуги из-за их низкого уровня качества, тогда бизнесу необходимо улучшать аспекты продукта до тех пор, пока пользователи не будут довольны. Во всех случаях предвидение оттока клиентов означает принятие упреждающих мер сейчас, а не позже, когда проблемы возникнут позже.

Другими словами, прогнозирование оттока помогает компаниям увеличить рентабельность инвестиций (ROI). Это также помогает повысить лояльность клиентов — если компания определила, что клиенты уходят из-за низкого качества или проблем с пользовательским интерфейсом, то компании необходимо улучшить эти аспекты, прежде чем ее пользователи уйдут в массовом порядке.

## **1.1.3 Причины использования машинного обучения для прогнозирования оттока**

Машинное обучение меняет многие аспекты нашей повседневной жизни, от рекомендации песен до оптимизации маршрутов путешествий.

Прогнозирование оттока — одно из самых известных применений машинного обучения, учитывая, что уровень оттока — решающий показатель для бизнеса. Использование МО для прогнозирования оттока может помочь понять и решить эту дорогостоящую проблему.

В описанном ранее пункте мы убедились, что прогнозировать отток клиентов важно для любой компании, и делать это нужно постоянно. Для человека, который бы не пользовался различными подручными средствами и машинным обучением, такая задача становится непосильной. Именно поэтому такой метод является самым оптимальным, быстрым и выгодным.

Подробнее весь процесс прогнозирования будет расписан во второй главе, где будет предоставлено описание всей практической части, но для общего понимания алгоритм МО в задачах предсказания оттока заключается в следующем: используются данные о клиентах, проводится их анализ, и модель прогнозирует вероятность оттока пользователей. Эти данные могут включать категориальные характеристики, такие как удовлетворенность клиентов и демографические данные, или числовую информацию, например их расходы. Таким образом, по завершению можно понимать группу лиц, которые в ближайшее время могут покинуть компанию, чтобы сотрудники имели возможность принять необходимые меры для их удержания.

Традиционно специалисты по науке о данных выполняли все эти шаги вручную, включая подготовку данных, предварительную обработку данных, выбор функций и реализацию классификаторов с помощью таких инструментов, как Python и sklearn, будь то логистическая регрессия, нейронные сети, глубокое обучение, случайный лес и т.п.

В курсовой работе также будет рассмотрен вопрос построения моделей ML для прогнозирования оттока клиентов. Для построения такой модели используются различные методы машинного обучения для решения задачи классификации. В результате нам удастся предсказывать вероятность того, что тот или иной человек захочет перестать пользоваться услугами, предоставляемые компанией, что в последующем может применяться и в реальной жизни.

# **1.2. Модели машинного обучения для решения задачи классификации**

Одна из типичных задач, возникающих в процессе определения поведения абонента — это определение принадлежности клиента к одному из двух классов: лояльных к компании и склонных к уходу. Этот процесс определения подразумевает задачу бинарной классификации.

Алгоритм классификации, основанный на обучающих данных, представляет собой метод контролируемого обучения, используемый для категоризации новых наблюдений. При классификации программа использует предоставленный набор данных или наблюдения, чтобы научиться классифицировать новые наблюдения по различным классам или группам. Например, 0 или 1, красный или синий, да или нет, спам или не спам и т. д. Для прогнозирования оттока клиентов в компании будет необходимо решать задачу бинарной классификации, где класс 0 будет означать, что клиент лоялен к компании и не собирается уходить, 1 – клиент склонен к уходу из компании или собирается перестать пользоваться предоставляемыми услугами.

Для решения этой задачи рассмотрим принцип работы 8 основных методов машинного обучения, используемых в задачах классификации.

## **1.2.1. Логистическая регрессия**

Логистическая регрессия — это соответствующий регрессионный анализ, который следует проводить, когда зависимая переменная имеет двоичное решение. Подобно всем другим типам регрессионных систем, логистическая регрессия также является типом системы прогностической регрессии. Логистическая регрессия используется для оценки связи между одной зависимой двоичной переменной и одной или несколькими независимыми переменными. Он дает дискретные выходы в диапазоне от 0 до 1.

Логистическая регрессия использует более сложную функцию затрат, чем линейная регрессия, эта функция затрат называется «сигмовидной функцией» вместо линейной функции.

Логистическая функция, также называемая сигмовидной функцией, первоначально использовалась статистиками для описания свойств роста популяции в экологии. Сигмовидная функция — математическая функция, используемая для сопоставления прогнозируемых значений с вероятностями. Логистическая регрессия может принимать значения от 0 до 1, но никогда точно в этих пределах. Формула для нее имеет следующий вид:

где f(x) = выход между 0 и 1 (оценка вероятности), x = вход в функцию.

Если построим это уравнение, то получим S-образную кривую логистической регрессии, как показано ниже.

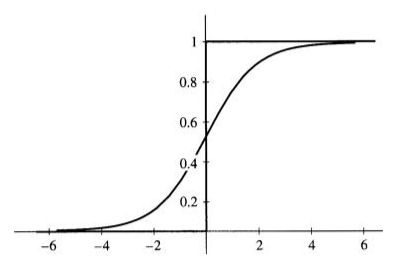


Рисунок 1. S-образная кривая логистической регрессии

Логистическая регрессия является продолжением модели линейной регрессии. Разберемся в этом на простом примере. Если мы хотим классифицировать, является ли электронное письмо спамом или нет, если мы применим модель линейной регрессии, мы получим только непрерывные значения от 0 до 1, такие как 0,4, 0,7 и т. Д. С другой стороны, логистическая регрессия расширяет эту модель линейной регрессии, устанавливая пороговое значение 0,5, поэтому точка данных будет классифицироваться как спам, если выходное значение больше 0,5, и не спам, если выходное значение меньше 0,5.

Таким образом, мы можем использовать логистическую регрессию для задач классификации и получать более точные прогнозы. Логистическая регрессия — это самый простой алгоритм бинарной классификации. Обычная регрессия в задачах классификации справляется значительно хуже из-за чувствительности и неограниченности. Именно поэтому, используя логистическую регрессию, результат которой часто интерпретируют как вероятность отнесения объекта к положительному классу, мы можем получать более точные прогнозы.

## **1.2.2. Метод опорных векторов**

Модель SVM или модель опорных векторов — это популярный набор моделей контролируемого обучения, которые используются для регрессии, а также классификационного анализа. Это модель, основанная на системе статистического обучения, и известна тем, что она надежна и эффективна в нескольких случаях использования. На основе невероятного двоичного линейного классификатора используется машина опорных векторов для разделения различных классов с помощью различных ядер.

Одна из основных причин, по которой компании склоняются к моделям машин с опорными векторами по сравнению с другими моделями, заключается в том, что машины опорных векторов имеют значительно более высокую точность, которую можно использовать при использовании уменьшенных вычислений из системы.

Метод опорных векторов находит лучший способ классификации данных на основе положения относительно границы между положительным классом и отрицательным классом. Эта граница известна как гиперплоскость, которая максимизирует расстояние между точками данных из разных классов.

Алгоритм линейной метода опорных векторов используется, когда у нас есть линейно разделяемые данные. Говоря простым языком, если у нас есть набор данных, который можно классифицировать на две группы с помощью простой прямой линии, мы называем его линейно разделяемыми данными, а классификатор, используемый для этого, известен как линейный классификатор SVM.

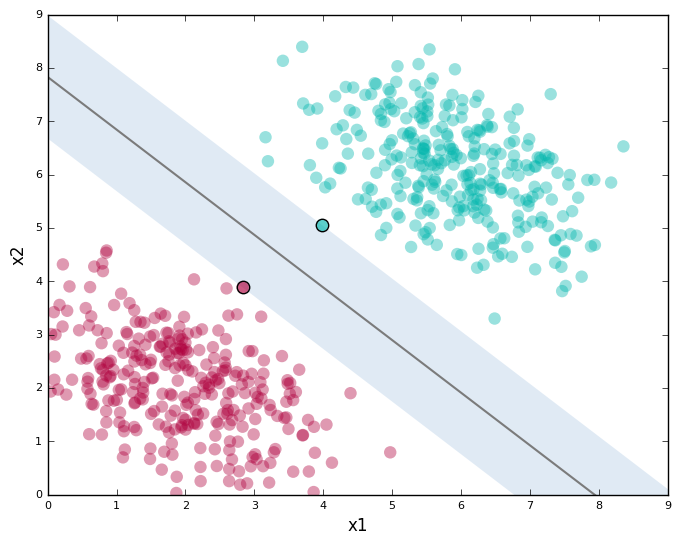


Рисунок 2. Линейный классификатор SVM

Граница принятия решения не обязательно должна быть линией. Нелинейный алгоритм метода опорных векторов используется, когда у нас есть нелинейно разделяемые данные. Подобно тому, как прямая может разделять точки в двумерном пространстве, гиперплоскость — это плоскость, которая разделяет точки в n-мерном пространстве.

Метод опорных векторов использует ядра, что являются функцией, на основе которой точки могут быть разделены. Точки, которые не являются линейно разделяемыми, являются проектами в пространство более высокой размерности.

Ядро в методе опорных векторов отвечает за преобразование входных данных в требуемый формат. Некоторые из ядер, используемых в SVM, являются линейными, полиномиальными и радиальными базисными функциями (RBF). Для создания нелинейной гиперплоскости мы используем RBF и полиномиальную функцию. Для сложных приложений следует использовать более продвинутые ядра для разделения нелинейных классов. С помощью этого преобразования можно получить точные классификаторы.

Таким образом, метод опорных векторов (SVM) — это метод, используемый как для классификации, так и для регрессии. Цель алгоритма SVM состоит в том, чтобы найти гиперплоскость в N-мерном пространстве, которое четко классифицирует точки данных. А возможность использовать различные ядра позволяет создавать лучшие модели, что в свою очередь помогает решать задачу классификации быстро и эффективно.

## **1.2.3. Алгоритм k-ближайших соседей (KNN)**

Алгоритм k-ближайших соседей (KNN)— это метод классификации данных для оценки вероятности того, что точка данных станет членом той или иной группы на основе того, к какой группе принадлежат ближайшие к ней точки данных.

Представим, что у нас есть три группы А (знак «+»), В (знак «-») и С (знак «∆»). Чтобы определить, к какой группе относится та или иная точка, алгоритм смотрит на состояния точек данных рядом с ней. Если большинство точек данных находятся в группе A, весьма вероятно, что рассматриваемая точка данных находится в группе A и наоборот. Именно поэтому, как и показано на рисунке, мы видим, что у предсказываемого нами объекта, из 5 соседей трое из них относятся к классу С, соответственно, и предсказываемый объект также относится к классу С.

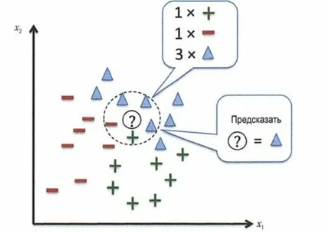


Рисунок 3. Принцип работы KNN

Другими словами, KNN включает в себя классификацию точки данных, глядя на ближайшую аннотированную точку данных, также известную как ближайший сосед.

Существует четыре способа расчета измерения расстояния между точкой данных и ее ближайшим соседом: евклидово расстояние, расстояние Манхэттена, расстояние Хэмминга и расстояние Минковского. Из них евклидово расстояние является наиболее часто используемой функцией расстояния или метрикой.

Таким образом, алгоритм k-ближайших соседей (KNN) — это простой алгоритм машинного обучения, который можно использовать для решения задач классификации. Его легко реализовать и понять, но имеет серьезный недостаток, заключающийся в зависимости от количества исследуемых объектов, а также в необходимости постоянного вычисления метрик расстояния между наблюдениями. Однако KNN по-прежнему может быть полезен при решении проблем, решение которых зависит от идентификации похожих объектов.

## **1.2.4 Дерево решений**

Классификаторы дерева решений работают как блок-схемы. Каждый узел дерева решений представляет собой точку принятия решения, которая разделяется на два листовых узла. Каждый из этих узлов представляет результат решения, и каждое из решений также может превратиться в узлы принятия решений. В конце концов, различные решения приведут к окончательной классификации.

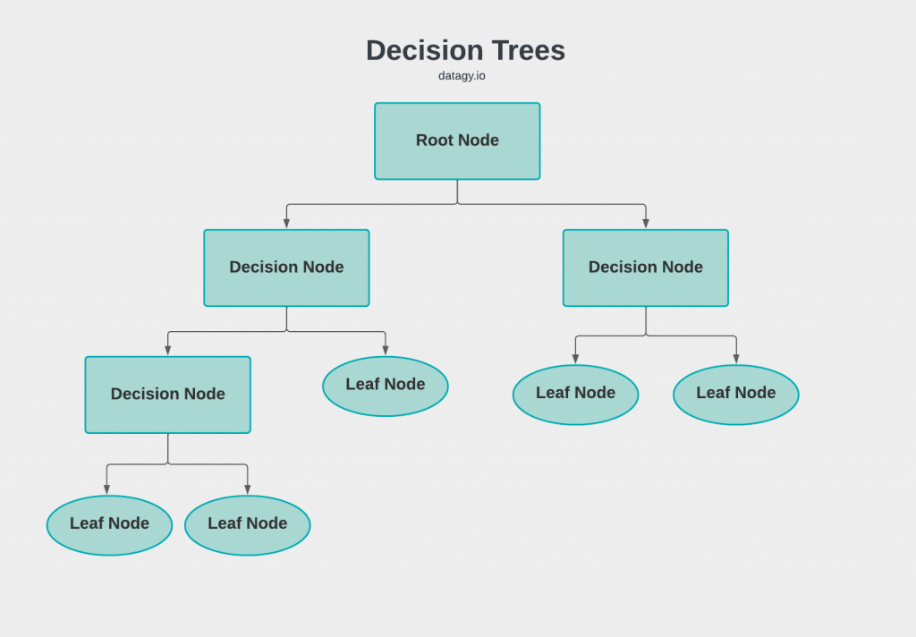
На приведенной ниже диаграмме показано, как деревья решений работают для принятия решений. Верхний узел называется корневым узлом. Каждая из точек принятия решений называется узлами принятия решений. Конечная точка принятия решения называется листовым узлом.

Рисунок 4. Визуализированный пример дерева решений

Идея алгоритма случайного леса состоит в том, чтобы случайным образом выбирать некоторые функции на узлах, возвращая их обратно, и выбирать лучшие функции в соответствии с требованиями учащегося. Чтобы разделить узлы на наиболее информативные функции, нам нужно определить целевую функцию, которую мы хотим оптимизировать с помощью алгоритма обучения дерева. Здесь наша целевая функция состоит в том, чтобы максимизировать получение информации при каждом разделении, которое мы определяем следующим образом:

где f- функция для выполнения разделения, Dp, Dleft и Dright являются набором данных родительского и дочернего узлов, I—мерой примеси, Np— общее число образцов на родительском узле, Nleft и Nright — количество выборок в дочерних узлах.

Когда строится модель дерева решений, то выделяют два этапа: индукция и отсечение. Первый этап заключается в построении дерева, учитывая исходные данные, чтобы установить границы решений. Второй же этап способен упростить модели, а также удалить ненужные структуры, что в свою очередь приводит к более понятному использованию алгоритма и его упрощению.

Дерево решений является одним из основных и широко используемых алгоритмов в области машинного обучения. Он используется в различных областях классификации и регрессионного моделирования. Благодаря своей способности отображать визуализированные выходные данные, можно легко извлечь идеи из процесса моделирования.

## **1.2.5. Случайный лес**

Random forest является одним из самых популярных алгоритмов машинного обучения на основе деревьев. Он также является наиболее гибким и простым в использовании.

Преимущества случайного леса многочисленны. Отдельные деревья принятия решений, как правило, чрезмерно соответствуют обучающим данным, но случайный лес может смягчить эту проблему, усреднив результаты прогнозирования с разных деревьев. Это дает случайным лесам более высокую прогностическую точность, чем одно дерево решений.

Алгоритм случайного леса работает, выполнив следующие шаги:

Шаг 1: Алгоритм выбирает случайные выборки из предоставленного набора данных.

Шаг 2: Алгоритм создаст дерево решений для каждого выбранного образца. Затем он получит результат прогнозирования из каждого созданного дерева решений.

Шаг 3: Затем будет проведено голосование за каждый прогнозируемый результат. Для задачи классификации он будет использовать моду, а для задачи регрессии – среднее значение.

Шаг 4: И, наконец, алгоритм выберет наиболее проголосованный результат прогноза в качестве окончательного прогноза.

Таким образом, случайный лес – один из эффективных методов решения в задачах классификации. Он преодолевает проблему перенастройки путем усреднения или объединения результатов различных деревьев решений. Random forest очень гибкий, обладает очень высокой точностью, а также имеет меньшую дисперсию, чем одно дерево решений. Именно поэтому во многих сферах, например, в медицине для выявления заболевания пациента или в банковском секторе на определение мошенников.

## **1.2.6. Наивный байесовский алгоритм**

Наивный байесовский классификатор — это вероятностная модель машинного обучения, которая используется для задачи классификации. Суть классификатора основана на теореме Байеса.

Используя теорему Байеса, мы можем найти вероятность возникновения А, учитывая, что произошло В. Здесь Б — доказательство, а А— гипотеза. Предположение, сделанное здесь, заключается в том, что предикторы /признаки независимы. То есть наличие одной конкретной особенности не влияет на другую. Поэтому его называют наивным.

Различают несколько типов наивного байсовского классификатора:

1. Мультиномиальный наивный Байес. Векторы признаков представляют частоты, с которыми определенные события были сгенерированы мультиномиальным распределением.
2. Бернулли наивный Байес. Применяется, когда признаки являются двоичными дискретными, т.е. могут принимать значения 0 и 1.
3. Гауссовский наивный Байес. Когда предикторы принимают непрерывное значение и не дискретны, предполагается, что эти значения выборки взяты из распределения Гаусса.

Наивный байесовский классификатор является простым и универсальным классификатором, который работает очень эффективно для больших наборов данных. С точки зрения производительности наивный байесовский классификатор имеет превосходную производительность по сравнению со многими другими классификаторами. Одним из главных недостатков наивного байесовского классификатора является присущее ему предположение о независимости между признаками. На практике функции в реальных наборах данных редко бывают независимыми. Несмотря на этот недостаток, наивный байесовский классификатор очень полезен при предварительном анализе и понимании структуры данных.

## **1.2.7. Градиентный бустинг**

Градиентный бустинг — это тип ускорения машинного обучения. Он опирается на интуицию, что наилучшая возможная следующая модель в сочетании с предыдущими моделями минимизирует общую ошибку прогнозирования. Ключевая идея состоит в том, чтобы установить целевые результаты для этой следующей модели, чтобы свести к минимуму ошибку. Целевой результат для каждого случая в данных зависит от того, насколько изменение прогноза этого случая влияет на общую ошибку прогнозирования:

* Если небольшое изменение прогноза для случая вызывает большое падение погрешности, то следующим целевым исходом дела является высокое значение. Прогнозы из новой модели, которые близки к ее целям, уменьшат ошибку.
* Если небольшое изменение прогноза для случая не приводит к изменению ошибки, то следующий целевой результат дела равен нулю. Изменение этого прогноза не уменьшает ошибку.

Градиентный бустинг возникает потому, что целевые результаты для каждого случая устанавливаются на основе градиента ошибки относительно прогноза. Каждая новая модель делает шаг в направлении, которое минимизирует ошибку прогнозирования, в пространстве возможных прогнозов для каждого обучающего случая.

# **1.3. Оценка моделей классификации**

В машинном обучения традиционно для выбора лучшей модели принято основываться на значения метрик. К основным метрикам относятся: accuracy, precision, recall, roc-auc и f1. Но для начала определимся с таким понятием, как матрица ошибок. Ее элементы представляют собой следующее:

1. TP - Истинные положительные результаты: истинное и предсказанное равны 1
2. TN - Истинные негативные результаты: истинная метка класса равна 0, и прогнозируемое также равно 0.
3. FN - Ложноотрицательный результат: истинное значение – 1, предсказанное – 0.
4. FP - Ложноположительный результат: истинное значение равно 0, а предсказанное – 1.

Не сложно догадаться, что в случае идеальной модели, в матрице ошибок будут ненулевые значения только на диагонали, которые отвечают за TN и TP, в то время как на местах FN и FP должны стоять нули. Но такие модели являются очень редким явлением, и в основном мы имеем ненулевые значения на местах FN и FP, которые также являются ошибками первого и второго рода соответственно.

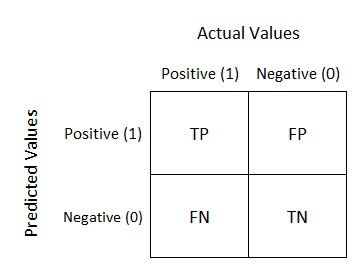


Рисунок 5. Матрица ошибок

Метрика accuracy - количество правильно поставленных результатов от общего количества данных. Эта метрика сильно зависит от баланса между классами, если данные несбалансированные, то ее применять не следует. Аccuracy вычисляется по формуле:

Метрика precision - количество истинно положительных метрик из всего набора положительных исходов, и вычисляется по формуле:

Метрика recall - это доля правильно классифицированных меток от общего количества положительных случаев.

F1 – это статистическая мера точности модели классификации. Её можно рассматривать как средневзвешенное значение точности и полноты модели, и она выражается формулой:

Кривая ROC используется для того, чтобы оценить способность предсказывать у модели. Эта кривая основывается на матрице путаницы (confusion matrix), где на оси абсцисс кривой отложена доля FP, а по оси ординат — доля TP.

AUC — это площадь под ROC-кривой. Чем больше AUC, тем сильнее способность модели идентифицировать положительные и отрицательные образцы. В таком случае, когда эта способность очень высокая, это позволяет получить лучший результат классификации, показывая производительность модели во время разложения классов.

# **Глава 2. Практическая часть**

# **2.1. Выбор набора данных для анализа**

Для выполнения практической части работы был выбран датасет на сайте Kaggle под названием «Telco Customer Churn». В нём предоставлена информация из анкетных данных клиентов и факт наличия оттока. Ниже представлено описание полей датасета:

* customerID – id клиента
* gender – пол клиента (male/female)
* SeniorCitizen – является ли клиент пенсионером (1, 0)
* Partner – состоит ли клиент в браке (Yes, No)
* Dependents – есть ли у клиента иждивенцы (Yes, No)
* tenure – сколько месяцев человек являлся клиентом компании
* PhoneService – подключена ли услуга телефонной связи (Yes, No)
* MultipleLines – подключены ли несколько телефонных линий (Yes, No, No phone service)
* InternetService – интернет-провайдер клиента (DSL, Fiber optic, No)
* OnlineSecurity – подключена ли услуга онлайн-безопасности (Yes, No, No internet service)
* OnlineBackup – подключена ли услуга online backup (Yes, No, No internet service)
* DeviceProtection – есть ли у клиента страховка оборудования (Yes, No, No internet service)
* TechSupport – подключена ли услуга технической поддержки (Yes, No, No internet service)
* StreamingTV – подключена ли услуга стримингового телевидения (Yes, No, No internet service)
* StreamingMovies – подключена ли услуга стримингового кинотеатра (Yes, No, No internet service)
* Contract – тип контракта клиента (Month-to-month, One year, Two year)
* PaperlessBilling – пользуется ли клиент безбумажным биллингом (Yes, No)
* PaymentMethod – метод оплаты (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
* MonthlyCharges – месячный размер оплаты на настоящий момент
* TotalCharges – общая сумма, которую клиент заплатил за услуги за все время
* Churn – произошел ли отток (Yes or No)

Решаемой задачей является построение наиболее эффективной модели машинного обучения, которая бы могла предсказывать вероятность прекращения использования услуг клиентом. На основе данных из анкет клиентов принимается решение о попытках удержать тех, кто находятся в группе предположительного оттока.

# **2.2. Предварительный анализ данных**

Первым делом загружаем данные из датасета в Jupyter Notebook, и в результате имеем следующую таблицу:

Изображение выглядит как текст, снимок экрана, внутренний

Автоматически созданное описание

Рисунок 6. Часть данных файла telecom.csv

В наборе данных отсутствуют дубликаты, а также пропущенные значения. Также, используя info(), можно узнать (Рисунок 7), что всего мы имеем по информацию о 7043 пользователях по 21 признаку.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 7. Общая информация по данным

Поскольку большинство признаков являются категориальными, то дальнейшее кодирование этих столбцов поможет провести дальнейший анализ. И в результате итоговый вид изначальных данных выглядит следующим образом:

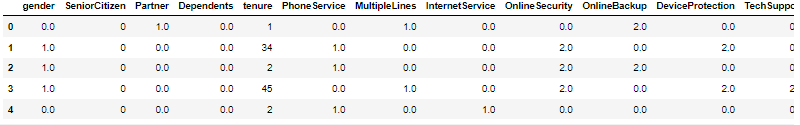


Рисунок 8. Закодированные категориальные признаки

Теперь можем приступать к анализу данных. Сначала, с помощью функции describe(), получим общую статистику по данным (Рисунок 9).

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 9. Общая статистика данных

Исходя из этого можно сделать следующие выводы:

* Медианный месячный платеж среди всех абонентов составляет 70.4 у.е.
* Большинство клиентов – мужчины,
* Медианное количество месяцев подписки составляет 29 месяцев. Максимальное – 72,
* Большинство клиентов не состоят в браке, а также не имеют иждивенцев,
* Интернет-провайдером большинства клиентов является Fiber optic,
* Большинство клиентов не пользуются дополнительными услугами,
* Больше половины клиентов имеют тип контракта Month-to-month,
* Большая часть клиентов использует безбумажный биллинг,
* Electronic check является приоритетным способом оплаты.

Также, посчитав общее количество ушедших людей и процент оттока на основе данных получаем, что всего из компании ушло 1886 клиентов, что в свою очередь составляет 27%.

Следующим шагом построим корреляционной матрицу и проверим мультиколлинеарность, чтобы понимать, какие признаки необходимо убрать. В результате проверки мультиколлинеарности, используя коэффициент ранговой корреляции Спирмена, получаем, что столбцы «TotalCharges» и «tenure» имеют высокую корреляцию. Это также доказывает корреляционная матрица (Рисунок 10).

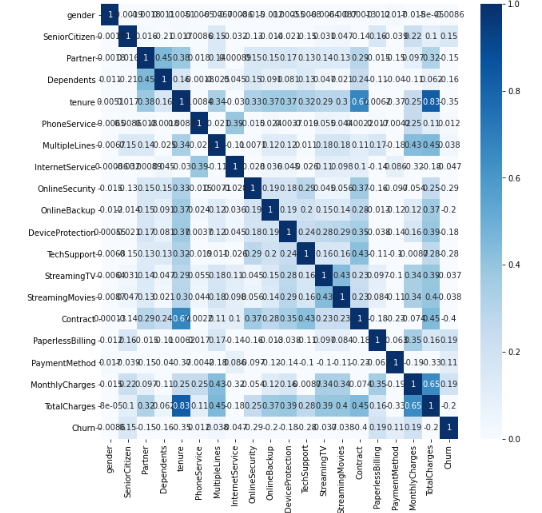


Рисунок 10. Корреляционная матрица

С целевой переменной Churn имеется слабая корреляционная связь у признаков: Tenure, OnlineSecurity, TechSupport, OnlineBackup, DeviceProtection, MonthlyCharges, PaperlessBilling. Наличие корреляционной связи этих признаков с целевой переменной говорит нам о том, что существует вероятность того, что клиент вероятно откажется от услуг связи, если у него не активированы услуги по онлайн безопасности, технической поддержке, защиты устройства и резервному копированию данных. Так же существует зависимость оттока и использования безбумажного биллинга. Зависимости между оттоком, количеством месяцев подписки и месячной платой являются следственными связями. Посмотрим сколько процентов клиентов пользовались услугами безопасности (Рисунок 11), а также выведем эти значения на график (Рисунок 12).

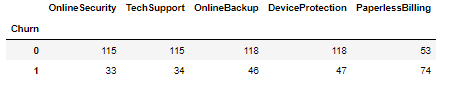
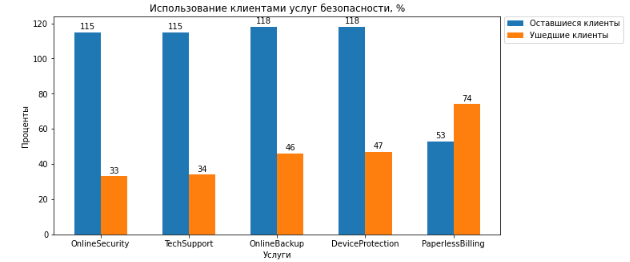


Рисунок 11. Процент использования услуг безопасности

Рисунок 12. Визуализация показателей использования услуг бесопасности

Особая разница заметна у соотношения клиентов услуг OnlineSecurity и TechSupport. Оставшиеся клиенты пользовались этими услугами в 3 раза чаще, а услуги OnlineBackup и DeviceProtection пользовались спросом у данной категории клиентов в почти 2 раза чаще. С безбумажным биллингом картина выглядит иначе - на 21% клиентов больше пользовались электронным счётом в группе оттока.

Теперь посмотрим на график зависимости процента ушедших клиентов от количества месяцев использования услуг (Рисунок 13). Исходя из него мы видим, что больше 20% клиентов ушли в первые месяцы пользования услугами. Учитывая то, как распределены дынные о месячной плате за связь среди 2 групп клиентов, то можно сделать вывод, что медианное значение группы оттока оказалось выше, чем значение у группы активных клиентов, несмотря на то что большинство ушедших клиентов не пользовалось услугами безопасности. Возможно, большинство этих клиентов посчитало, что переплачивает за услуги.

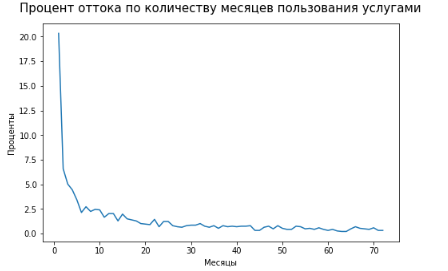


Рисунок 13. График % оттока по количеству месяцев пользования услугами

Теперь перейдем к анализу непрерывных данных. Построив гистограммы распределений (Рисунок 14) и ящики с усами (Рисунок 15), нельзя сделать вывод о том, что какой-то из этих признаков совершенно не влияет на отток, поэтому их мы не исключаем. По ним можем сделать заключение, что оттоку подвержены те, кто имеют большую месячную плату, но несмотря на это, те, кто больше всего заплатил за все время за услуги, остались клиентами компании. Возможно, это из-за того, что большая часть этих людей являются постоянными клиентами.

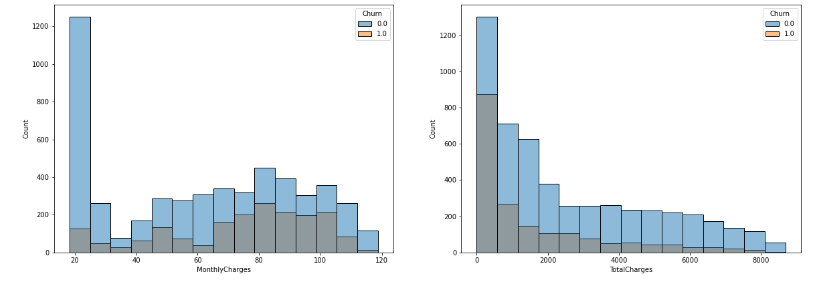


Рисунок 14. Гистограммы распределений

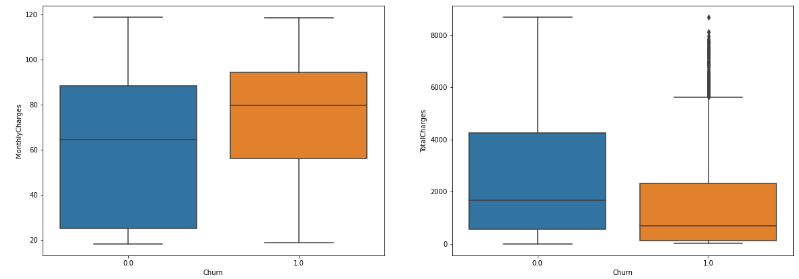


Рисунок 15. Ящики с усами

Переходя к анализу категориальных признаков (Рисунок 16), мы видим, что признак, который совсем не оказывал влияния на распределение смертей – gender. Также распределения MultipleLines и StreamingTV для класса 0 и класса 1 были похожими, что означало, что характеристики признаков не отличаются в зависимости от метки клиента.

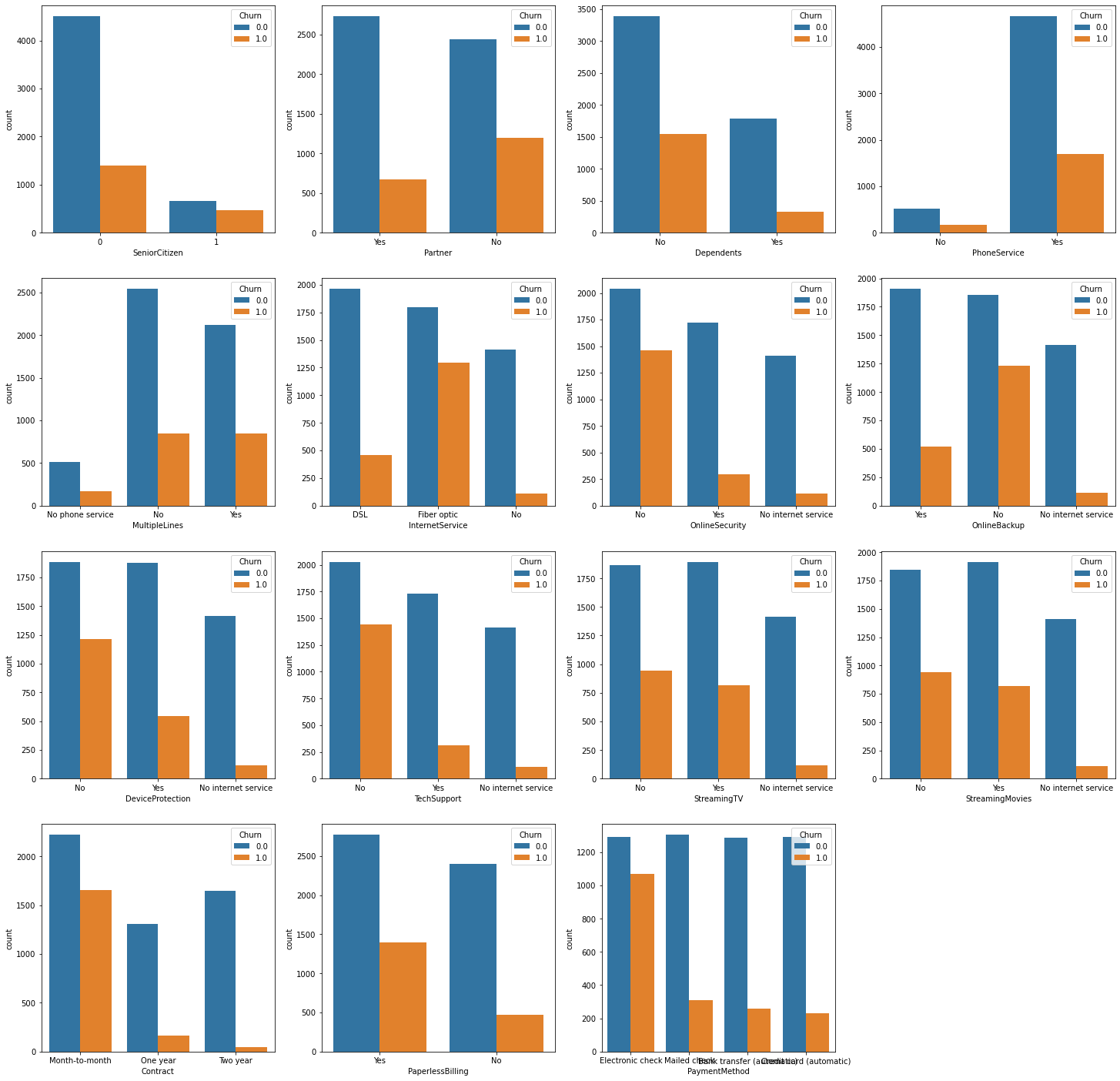


Рисунок 16. Категориальные данные

Также для наглядности были выведенные сводные графики по каждому из признаков для понимания общей картины происходящего влияния параметра на отток (Рисунок 17).

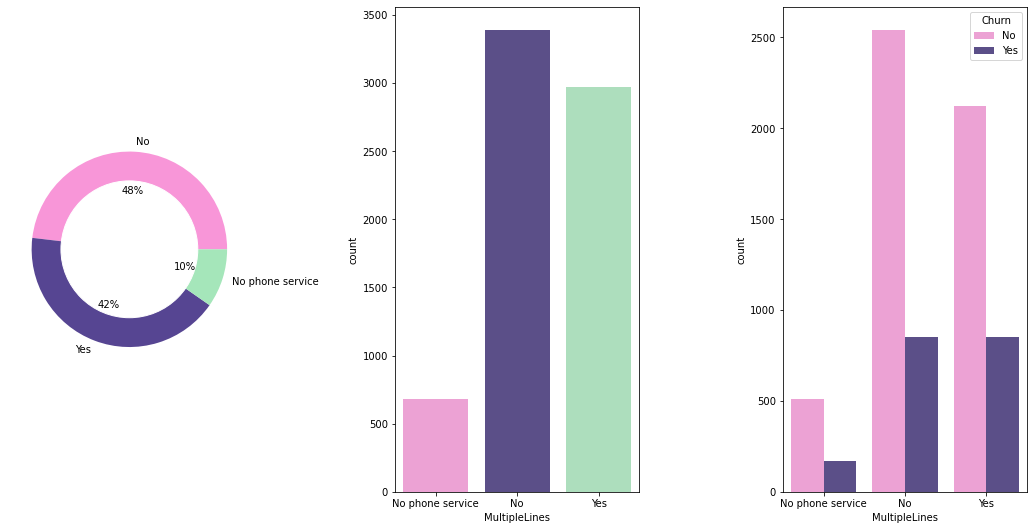


Рисунок 17. Статистика по категориальному признаку

Таким образом, получаем, одной из основных причин оттока клиентов является отсутствие подписки на услуги безопасности и использование электронных счетов. По причине отсутствия подписок на услуги безопасности, ушедшие клиенты переплачивали за месячное пользование связью. Около 20% ушедших клиентов пользовалась услугами около месяца. Больше половины клиентов из группы оттока ушли пользуясь связью менее года. В результате получаем, что из данных необходимо убрать столбцы tenure, gender, MultipleLines, StreamingTV и customerID, значение которого никак не влияет на отток. На этом предобработка данных заканчивается. В итоге набор данных содержит 17 независимых переменных и целевую переменную churn.

Перед построением моделей было совершено понижение размерности до 2-х компонент (Рисунок 18). Как видно из рисунка, центр красных точек, отвечающих за отток, смещен относительно центра синих (оставшихся клиентов) левее, и большая часть точек располагается в той же области.

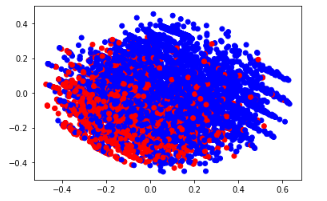


Рисунок 18. Понижение размерности

# **2.3. Построение моделей**

Для построения моделей разобьем нашу выборку на обучающую и тестовую в отношении 80% и 20% соответственно. Но перед этим проведем стандартизацию данных для устранения большой размерности и получения более качественного прогноза.

Построим следующие 11 моделей с использованием кросс-валидации, а именно:

1. Логистическая регрессия (Logistic Regression)
2. Дерево решений (Decision Tree Classifier)
3. Случайный лес (Random Forest Classifier)
4. Метод k-ближайших соседей (KNN)
5. Метод опорных векторов с ядром linear (SVM – linear)
6. Метод опорных векторов с ядром poly (SVM – poly)
7. Метод опорных векторов с ядром rbf (SVM – rbf)
8. Метод опорных векторов с ядром sigmoid (SVM – sigmoid)
9. XGBoost
10. Наивный байессовский метод (GaussianNB)
11. Градиентный бустинг (GradientBoostingClassifier)

В результате получаем следующую таблицу по моделям:

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 19. Отчет о моделях

В этой таблице столбцы обозначают следующее:

* модель – название модели
* roc\_auc, accuracy, precision, recall, f1 – значения метрик после обучения на конкретном разделении
* cross valid – значение f1\_score при проведении кросс-валидации
* time – время, которое было потрачено на процесс обучения модели

В результате можно выделить следующие модели:

* RandomForestClassifier обладает самым высоким значением ROC-AUC и precision
* LogisticRegression имеет высокие показатели по всем метрикам, что также является поводом для выбора этой модели
* GaussianNB - самая высокая точность на кросс-валидации, а также по матрице неточностей можно заметить, что данная модель чаще предсказывает отток, но значение ROC-AUC наименьшее

Теперь с помощью GridSearch можем сделать данные модели лучше для более точного прогноза, подоврав такие параметры для моделей, которые будут лучшими. В результате имеем:

* для LogisticRegression значение регуляризации С – 0,001 и решатель solver – liblinear.
* для RandomForestClassifier criterion – gini – и максимальная глубина дерева – 6.
* для GaussianNB - var\_smoothing = 0.01

Помимо GridSearchCV для усовершенствования моделей добавим суррогатные признаки. Для этого будем использовать метод анализа главных компонент (PCA). Уменьшим размерность признаков до указанного количества компонент с сохранением информации, в нашем случае 5. Чтобы построить модели, которые будут лучше предыдущих, то необходимо добавить суррогатные признаки к изначальным, и снова строить новые модели. Видим, что после применения grid search все модели по всем метрикам гораздо улучшились. ROC-AUC везде превышает 0.82, Precision самое большое вышло у RandomForest, а значение метрики cross\_val наибольшее у наивного байессовского метода. В итоге можно сказать, что , а введение суррогатных признаков хорошо сказывается на качестве предсказания.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 20. Полный отчет обо всех моделях

Но на этом варианты улучшения моделей не заканчиваются, и следующим шагом будет использование бэггинга, в котором за основной классификатор будет браться GaussianNB, а затем LogisticRegression с гиперпараметрами, которые были определены ранее. Помимо этого, будем использовать суррогатные признаки, в результате чего имеем новые три модели - Bagging LogisticRegression, Bagging LogisticRegression PCA и Bagging GaussianNB (Рисунок 20). В результате это не дало никаких улучшений на кросс-валидации, но при этом при использовании бэггинга немного улучшились метрики roc\_auc, accuracy и precision.

После этого попробуем создать голосующий ансамбль на основе ранее создаваемых функций. Голосующий ансамблю будет состоять из трех моделей, полученные ранее после подбора гиперпараметров, из двух – наивного байесовского и логистической регрессии с применением суррогатных признаков, и из двух - логистической регрессии с применением суррогатных признаков и случайный лес с применением гиперпараметров. В итоге получаем три новые модели - VotingClassifier-3, VotingClassifie-2-PCA и VotingClassifie-2-Log\_RF\_PCA. По данным из таблицы можно сделать вывод, что такими действиями мы достигли увеличения метрик кросс валидации и f1-меры, которые стали самыми максимальными среди всех моделей, а также улучшили показатели roc-auc.

Таким образом, модели Random Forest и Logistic Regression имеют самые высокие результаты, также они не занимают много времени для обучения. При этом, если их объединить в голосующий ансамбль, то надежность повышается, а показатели roc-auc и f1-меры становятся наибольшими, что в свою очередь приводит к выбору именно этих моделей.

# **ЗАКЛЮЧЕНИЕ**

В настоящее время компании сталкивается с жесткой конкуренцией. Прогнозирование оттока клиентов стало важной задачей управления взаимоотношениями с клиентами для удержания ценных клиентов. Поэтому, проведя исследование, можно будет хорошо понять ключевые факторы оттока для удержания людей и их влияние на отток.

Правильное управление оттоком может сэкономить огромную сумму денег для компании. Таким образом, экономическую ценность удержания клиентов можно резюмировать следующим образом:

* довольные клиенты могут привлечь новых клиентов
* долгосрочные клиенты обычно не испытывают сильного влияния конкурентов
* долгосрочные клиенты, как правило, покупают больше
* компания может сосредоточиться на удовлетворении потребностей существующих клиентов
* потерянные клиенты делятся негативным опытом и, таким образом, окажут негативное влияние на имидж компании

Таким образом понимаю группу лиц, находящихся в группе тех, кто с большей вероятность прекратят пользоваться услугами, компании могут их сохранив, тем самым не потерять не только ценные кадры, но и сохранить свою компанию в целом.

В курсовой работе была рассмотрена именно проблема оттока клиентов из компании, которая будет важная для любого бизнеса всегда. Благодаря машинному обучению удалось создать 22 модели для прогноза оттока клиентов. Для этого было использовано 7 различных алгоритмов, методов улучшения моделей, а также создания более сложных структур.

В результате с применением логистической регрессии и внедрения суррогатных признаков удалось получить самый качественный прогноз из всех представленных. Такую модель можно применять в больших компаниях и не только для отслеживания поведения клиентов и определения тех людей, которые могут покинуть компанию в ближайшее время для своевременного исправления такой ситуации.

# **СПИСОК ЛИТЕРАТУРЫ**

1. 5 Types of Classification Algorithms in Machine Learning // MonkeyLearn // URL: https://monkeylearn.com/blog/classification-algorithms/ (дата обращения: 04.12.2022).
2. 7 Types of Classification Algorithms in Machine Learning // URL: https://www.projectpro.io/article/7-types-of-classification-algorithms-in-machine-learning/435 (дата обращения: 04.12.2022).
3. A Tutorial for Churn Prediction Using Machine Learning// Business Intelligence // URL: https://www.akkio.com/post/a-tutorial-for-churn-prediction-using-machine-learning (дата обращения: 04.12.2022).
4. Customer Churn Prediction Using Machine Learning: Main Approaches and Models // URL: https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html (дата обращения: 04.12.2022).
5. Classification in Machine Learning: What it is and Classification Models // URL: https://www.simplilearn.com/tutorials/machine-learning-tutorial/classification-in-machine-learning (дата обращения: 04.12.2022).
6. Gradient Boosting In Classification // URL: https://blog.paperspace.com/gradient-boosting-for-classification/ (дата обращения: 04.12.2022).
7. How to Implement Customer Churn Prediction [Machine Learning Guide for Programmers] // MLOps Blog // URL: https://neptune.ai/blog/how-to-implement-customer-churn-prediction (дата обращения: 04.12.2022).
8. How to Predict Customer Churn with Machine Learning // URL: https://www.phdata.io/blog/determining-customer-churn-with-ml/ (дата обращения: 04.12.2022).
9. Introduction to decision tree classifiers from scikit-learn // URL: https://towardsdatascience.com/introduction-to-decision-tree-classifiers-from-scikit-learn-32cd5d23f4d (дата обращения: 04.12.2022).
10. Logistic Regression in Machine Learning // URL: https://www.scaler.com/topics/machine-learning/logistic-regression-machine-learning/ (дата обращения: 04.12.2022).
11. Support Vector Machines in Machine Learning (SVM): 2023 Guide // URL: https://www.knowledgehut.com/blog/data-science/support-vector-machines-in-machine-learning (дата обращения: 04.12.2022).
12. Top 6 Machine Learning Algorithms for Classification // URL: https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501 (дата обращения: 04.12.2022).
13. А. А. Карякина, А. В. Мельников, Сравнение моделей прогнозирования оттока клиентов интернет-провайдеров // URL: http://jmlda.org/papers/doc/2017/no4/Karyakina2017Churn.pdf (дата обращения: 04.12.2022).
14. Построение модели оттока сотрудников для разработки стратегии удержания // Хабр // URL: https://habr.com/ru/post/582304/ (дата обращения: 04.12.2022).
15. Предсказание оттока абонентов: Сравнение методов машинного обучения // Компьютерные инструменты в образовании, 2018 № 5: 5–23 (дата обращения: 04.12.2022).
16. Прогнозируйте отток клиентов с помощью Python и машинного обучения // IT CHEF// URL: https://itchef.ru/articles/254364/ (дата обращения: 04.12.2022).
17. Телекоммуникационные отрасли // Использование машинного обучения для прогноза оттока в телекоммуникационной отрасли // URL: https://www.internauka.org/journal/science/internauka/238/337259 (дата обращения: 04.12.2022).

# **ПРИЛОЖЕНИЕ**

Код программы доступен по ссылке:

https://github.com/Julia-491/Customer-Churn-Filmonova-PM19-3