Geekbrains

**Прогнозирование оттока клиентов в телекоммуникационной компании с применением алгоритмов машинного обучения**

Программа: Разработчик-аналитик

Специализация Data Engineer

Колычев Алексей Валерьянович

Москва

2024

**СОДЕРЖАНИЕ**

Введение……………………………………………………………………………...2

Глава 1. Основы машинного обучения: виды и их алгоритмы…………................4

1.1 Что такое машинное обучение………………...………………………………...4

* 1. Основные типы и виды машинного обучения………………………………….5
  2. Основные алгоритмы машинного обучения…………………………………...6
  3. Инструменты в машинном обучении………………………………………….12

..

Глава 2. Проведение анализа данных компании в Jupiter notebook……………..15

2.1 Исследовательский анализ данных в Jupiter notebook. Часть 1………………15

2.2 Исследовательский анализ данных в Jupiter notebook. Часть 2………………25

Глава 3. Применение моделей машинного обучения для прогнозирования оттока клиентов……………………………………………………………………………..30

3.1 Выбор модели машинного обучения…………………………………………..30

3.2 Модель Decision Tree…………………………………………………………...32

3.3 Модель Random Forest………………………………………………………….36

3.4 Модель Boosted Trees…………………………………………………………...38

3.5 Модель логистической регрессии……………………………………………...41

3.6 Метод опорных векторов (SVM)……………………………………………….45

3.7 Метод K-ближайших соседей (KNN)………………………………………….46

3.8 Общий вывод……………………………………………………………………48

Заключение………………………………………………………………………….49

**Введение**

В настоящее время конкуренция на рынке телекоммуникационных услуг достигла своего пика, и удержание клиентов стало одним из ключевых элементов успешной стратегии компании. Одной из основных проблем, с которой сталкиваются телекоммуникационные компании, является отток клиентов. Потеря клиентов может оказать негативное влияние на финансовые показатели компании, а также на её репутацию.

Целью данного проекта является прогнозирование оттока клиентов в телекоммуникационной компании с применением алгоритмов машинного обучения и разработка рекомендаций по их удержанию. Данный проект будет решать проблему определения клиентов, которые могут уйти от компании, и выявлять ключевые факторы, влияющие на решение клиента остаться или уйти.

Для достижения поставленной цели будут использоваться алгоритмы машинного обучения, такие как деревья принятия решений, случайный лес и градиентный бустинг, модель логистической регрессии, метод опорных векторов (SVM), метод K-ближайших соседей (KNN). Эти алгоритмы позволят проанализировать исторические данные о клиентах компании, идентифицировать паттерны и предсказать вероятность оттока для каждого клиента.

При проведении исследования использовались такие инструменты как Python, Jupiter notebook и различные библиотеки для анализа данных.

В качестве исторических данных были взяты данные компании с сайта [www.kaggle.ru](http://www.kaggle.ru), где можно найти много исторических данных различных компаний.

Применение машинного обучения в данном проекте позволит компании оптимизировать свои маркетинговые и удерживающие стратегии, направив усилия на клиентов с наибольшей вероятностью оттока. Благодаря этому компания сможет сохранить большее количество клиентов и увеличить свою доходность.

Актуальность данной темы обусловлена необходимостью телекоммуникационным компаниям улучшить качество обслуживания клиентов, минимизировать отток и увеличить свою конкурентоспособность на рынке. Применение алгоритмов машинного обучения в данном проекте позволит компании совершить качественный прорыв в сфере удержания клиентов и повысить свою эффективность.

**Глава 1. Основы машинного обучения: виды и их алгоритмы**

**1.1 Что такое машинное обучение**

Машинное обучение – это раздел искусственного интеллекта, который изучает методы, алгоритмы и модели, позволяющие компьютеру обучаться на основе данных, не явно программируясь на определенные инструкции. В отличие от классического программирования, где набор инструкций задается заранее, в машинном обучении компьютер самостоятельно определяет закономерности и шаблоны в данных, чтобы делать прогнозы или принимать решения.

Используя технологию машинного обучения, программисты больше не обязаны тратить время на написание инструкций, рассматривающих все возможные сценарии и содержащих все решения. Вместо этого они могут встроить в компьютер или программу алгоритм, который самостоятельно находит решения, используя статистические данные для выявления закономерностей и предсказаний.

Технология машинного обучения на основе анализа данных впервые появилась в 1950-х годах при разработке программ для игры в шашки. За прошедшие десятилетия этот общий принцип остался неизменным, но благодаря резкому увеличению вычислительной мощности компьютеров, стало возможным создание более сложных закономерностей и прогнозов, а также решение более широкого круга задач с применением машинного обучения.

Для начала процесса машинного обучения необходимо загрузить в компьютер набор данных, на которых алгоритм будет учиться. Например, это могут быть изображения собак и кошек с уже расставленными метками, указывающими на их вид. После обучения программа сможет самостоятельно распознавать собак и кошек на новых фотографиях без меток. Чем больше данных обработала программа, тем точнее будет ее распознавание.

Сегодня благодаря машинному обучению компьютеры могут распознавать и фотографии, и изображения, и тексты. Например, программы уже умеют распознавать не только лица на фотографиях, но и пейзажи, предметы, текст и цифры. Для текстов также необходимо машинное обучение: проверка грамматики в текстовых редакторах и на мобильных устройствах уже стала стандартом. Также существуют программы, способные автоматически создавать новостные статьи на различные темы без участия человека.

**1.2 Основные типы и виды машинного обучения**

Все задачи, решаемые с помощью ML, относятся к одной из следующих категорий:

1) *Задача регрессии* заключается в предсказании числового значения на основе данных с различными признаками. Например, прогнозирование цены акций через определенный период времени или ожидаемый объем продаж товара на следующий месяц.

2) *Задача классификации* состоит в присвоении объекту определенной категории на основе набора признаков. Например, определение наличия определенного объекта на изображении или диагностика болезни по медицинским показателям.

3) *Задача кластеризации* предполагает разделение данных на группы по их сходству без предварительного определения категорий. Например, классификация покупателей по их покупательским привычкам или разделение текстов по тематике.

4) *Задача уменьшения* размерности заключается в сокращении количества признаков для упрощения анализа данных и улучшения визуализации. Например, сокращение размерности для отображения данных на двумерном графике.

5) *Задача выявления* аномалий заключается в обнаружении нестандартных и редких случаев в данных. Например, выявление мошеннических операций с кредитными картами или выявление необычных поведенческих шаблонов в данных.

*Основные виды машинного обучения*

Машинное обучение подразделяется на два основных типа: обучение с учителем и обучение без учителя. В обоих случаях машинам предоставляются исходные данные для анализа и выявления закономерностей, но различие заключается в участии человека в обучении.

1. *Обучение с учителем*

При обучении с учителем машине предоставляются данные вместе с ответами, что позволяет ей проверять свои гипотезы. Задача заключается в том, чтобы создать модель, которая может предсказывать целевую переменную (например, цену квартиры) на основе входных данных (площадь, расположение и т. д.). Примеры задач обучения с учителем включают регрессию (предсказание непрерывных значений, таких как цена) и классификацию (разделение объектов на категории, такие как спам или не спам).

1. *Обучение без учителя*

В обучении без учителя машине не предоставляются ответы. Она должна самостоятельно обнаруживать шаблоны и структуры в данных. Задачи обучения без учителя включают:

- Кластеризация: Разделение данных на группы на основе сходства, например, в случае распределения людей по размерным группам для пошива рубашек.

- Уменьшение размерности: Уменьшение количества признаков в данных для облегчения их отображения и анализа. Например, отображение данных с сотнями признаков в двух- или трехмерном пространстве.

**1.3 Основные алгоритмы машинного обучения**

Алгоритмы моделей машинного обучения – это математические процедуры, которые используются для обучения моделей машинного обучения на данных. Они определяют, как модель будет учиться и делать прогнозы. Существует множество различных алгоритмов машинного обучения, каждый со своими сильными и слабыми сторонами.

Вот некоторые общие категории алгоритмов машинного обучения.

1. *Дерево принятия решений для бизнеса*

Дерево принятия решений – это метод, который помогает принимать решения с учетом потенциальных последствий, эффективности и затрат. В контексте бизнес-процессов дерево принятия решений создается путем последовательности вопросов "да/нет", которые ведут к правильному выбору.

Этот метод структурирует и систематизирует проблему, обеспечивая логическую основу для принятия решений. Дерево принятия решений особенно полезно, когда необходимо учитывать несколько факторов и оценивать вероятность различных событий.

Вот как работает дерево принятия решений для бизнеса:

- Начните с определения проблемы или решения, которое нужно принять.

- Создайте корневой узел, представляющий исходную проблему.

- Добавьте ветви для каждого возможного решения или действия.

- Для каждой ветви добавьте узлы, представляющие возможные последствия, эффективность и затраты.

- Повторяйте этот процесс, создавая дочерние узлы, пока не достигнете конечных узлов, представляющих конкретные решения.

- Оцените вероятность и влияние каждого возможного пути.

- Выберите путь, который приводит к наилучшему результату с учетом всех факторов.

Дерево принятия решений помогает предприятиям принимать обоснованные решения, учитывая все релевантные факторы и возможные последствия. Оно обеспечивает прозрачный и систематический подход к принятию решений, что особенно ценно в сложных и неопределенных ситуациях.

1. *Наивная байесовская классификация*

Наивная байесовская классификация – это метод машинного обучения, используемый для классификации данных. Он основан на теореме Байеса, которая позволяет рассчитывать вероятность события на основе имеющейся информации.

Наивность в названии метода заключается в предположении, что признаки независимы друг от друга, даже если на практике это может быть и не так. Несмотря на это упрощение, наивные байесовские классификаторы часто показывают хорошие результаты в реальных задачах.

Вот как работает наивная байесовская классификация:

- Рассчитать вероятность того, что объект принадлежит каждому классу на основе имеющихся данных.

- Рассчитать вероятность того, что объект имеет конкретный набор признаков для каждого класса.

- Умножить вероятности из шагов 1 и 2 для каждого класса.

- Классифицировать объект в класс с наибольшим произведением вероятностей.

Наивные байесовские классификаторы широко используются в следующих областях:

- Фильтрация спама

- Классификация новостных статей

- Анализ настроений

- Распознавание лиц и образов

Они популярны благодаря своей простоте, эффективности и возможности работать с большим количеством признаков и данных.

1. *Метод наименьших квадратов*

Метод наименьших квадратов ­– это статистический метод, используемый для подгонки прямой к набору данных. Цель метода - найти прямую, которая наиболее точно соответствует точкам данных, минимизируя сумму квадратов расстояний между точками и прямой.

Как работает метод наименьших квадратов:

- Предварительная обработка данных: данные преобразуются в числовой формат, и для каждого объекта создается набор признаков.

- Определение целевой переменной: выбирается целевая переменная, которую необходимо предсказать.

- Подгонка прямой: алгоритм метода наименьших квадратов используется для подгонки прямой к данным. Алгоритм находит значения параметров прямой, которые минимизируют сумму квадратов расстояний между точками данных и прямой.

- Прогнозирование: после подгонки прямой ее можно использовать для прогнозирования целевой переменной для новых объектов.

Метод наименьших квадратов широко используется в машинном обучении для решения задач регрессии. Регрессия – это задача прогнозирования непрерывных целевых переменных (например, цены на акции или продаж).

Преимущества:

- Простой и эффективный алгоритм

- Хорошо подходит для задач регрессии с линейными зависимостями

- Может работать с большим количеством признаков и данных

Недостатки:

- Предположение о линейной зависимости может снизить точность в некоторых случаях

- Может быть чувствителен к шуму и выбросам в данных

Области применения:

- Прогнозирование спроса

- Анализ временных рядов

- Оценка рисков

- Финансовое моделирование

*4) Логистическая регрессия*

Логистическая регрессия – это метод анализа данных, который используется для предсказания вероятности возникновения определенных событий на основе одной или нескольких независимых переменных. Она применяется в различных областях, таких как кредитный скоринг, анализ рекламных кампаний, прогнозирование прибыли от продаж товаров и предсказание естественных катастроф. Логистическая регрессия позволяет оценить влияние различных факторов на вероятность наступления события и принять обоснованные решения на его основе.

1. *Метод опорных векторов*

Метод опорных векторов (SVM) представляет собой набор алгоритмов, которые широко используются для решения задач классификации и регрессионного анализа. SVM строит гиперплоскость в N-мерном пространстве, чтобы разделить объекты на два класса. Гиперплоскость создается таким образом, чтобы максимально удалиться от ближайшей точки каждого класса.

Этот метод помогает решать разнообразные задачи машинного обучения, такие как сплайсинг ДНК, определение пола человека по фотографии, и показ рекламы на сайтах.

1. *Метод ансамблей*

Метод ансамблей в машинном обучении – это подход, при котором используется не один классификатор, а множество классификаторов, которые работают вместе для принятия решения. Этот метод снижает влияние случайных ошибок и уменьшает дисперсию результатов, так как объединение нескольких моделей дает более точные прогнозы, чем одна отдельно взятая.

Подход ансамбль учитывает различные гипотезы и позволяет расширить множество базовых гипотез для более точного прогнозирования.

1. *Кластеризация*

Кластеризация – это процесс разделения группы объектов на подмножества, называемые кластерами, таким образом, чтобы объекты в каждом кластере были более похожи друг на друга, чем на объекты из других кластеров.

Существует несколько алгоритмов для кластеризации объектов. Некоторые из них включают:

- Алгоритмы на основе центра тяжести, которые определяют кластеры на основе расстояния между объектами и их центром тяжести.

- Алгоритмы подключения, которые определяют кластеры на основе связей между объектами.

- Алгоритмы сокращения размерности, которые уменьшают размерность данных для упрощения кластеризации.

- Алгоритмы плотности, которые опираются на плотность объектов в пространстве для определения кластеров.

- Вероятностные алгоритмы, которые используют вероятностные модели для кластеризации.

- Методы машинного обучения, включая нейронные сети, для кластеризации данных.

Алгоритмы кластеризации применяются в различных областях, таких как биология (для анализа геномов и генетических данных), социология (для анализа социологических данных) и информационные технологии (для группировки данных и поиска закономерностей).

1. *Метод главных компонент*

Метод главных компонент (PCA) – это статистический метод ортогонального преобразования, который используется для преобразования наблюдений над переменными, имеющими какие-то взаимосвязи, в новый набор главных компонентов – линейно некоррелированных значений.

PCA обычно применяется для визуализации данных и сжатия информации, что позволяет упростить и уменьшить размерность данных для дальнейшего анализа. Однако PCA не подходит для данных, в которых все компоненты имеют высокую дисперсию и слабо упорядочены. Поэтому его применимость зависит от того, насколько хорошо изучена предметная область и какие данные доступны для анализа.

1. *Сингулярное разложение*

Сингулярное разложение, или SVD, в линейной алгебре представляет собой разложение прямоугольной матрицы на произведение трех матриц: U, Σ и V, где U и V являются унитарными матрицами, а Σ - диагональной матрицей.

Метод главных компонент является частным случаем сингулярного разложения. В прошлом алгоритмы компьютерного зрения основывались на этом методе, представляя лица или другие объекты как сумму базисных компонент, уменьшая их размерность и сравнивая с образцами из выборки. Современные алгоритмы сингулярного разложения в машинном обучении намного сложнее своих предшественников, но их основной принцип остается неизменным.

1. *Анализ независимых компонент*

Анализ независимых компонент, или ICA – это метод, который позволяет выявить скрытые факторы, влияющие на случайные величины или сигналы. Он строит модель, объясняющую данные с использованием независимых компонентов, которые считаются негауссовскими сигналами. В отличие от анализа главных компонент, ICA эффективнее в ситуациях, когда стандартные методы неэффективны. Этот метод широко используется в различных областях, таких как астрономия, медицина, распознавание речи, финансовый анализ и другие. Он помогает обнаружить скрытые причины явлений и является мощным инструментом для исследований и анализа данных.

**1.4 Инструменты в машинном обучении**

Инструменты машинного обучения – это программные платформы или библиотеки, которые предоставляют набор функций и алгоритмов для разработки, обучения и развертывания моделей машинного обучения. Эти инструменты позволяют разработчикам и специалистам по данным автоматизировать процессы машинного обучения и создавать более точные и эффективные модели.

Вот некоторые из наиболее популярных инструментов машинного обучения:

- TensorFlow: открытая платформа с открытым исходным кодом, разработанная Google. TensorFlow предоставляет широкий спектр инструментов и библиотек для создания, обучения и развертывания моделей глубокого обучения и других типов моделей машинного обучения.

- scikit-learn: библиотека Python с открытым исходным кодом, предоставляющая широкий спектр алгоритмов машинного обучения для классификации, регрессии, кластеризации и других задач.

- PyTorch: библиотека Python с открытым исходным кодом, разработанная Facebook. PyTorch используется в основном для глубокого обучения и предоставляет гибкую и эффективную среду для создания и обучения моделей.

- Keras: высокоуровневый API для TensorFlow, разработанный для упрощения создания и обучения моделей глубокого обучения. Keras предоставляет простой и удобный интерфейс для создания и запуска моделей с использованием TensorFlow.

- Jupyter Notebook: веб-приложение с открытым исходным кодом, которое позволяет пользователям создавать и делиться документами, содержащими живой код, уравнения, визуализации и текстовые пояснения. Jupyter Notebook широко используется для исследований в области науки о данных и разработки моделей машинного обучения.

Другие популярные инструменты машинного обучения:

- XGBoost: библиотека с открытым исходным кодом для градиентного бустинга, используемая для классификации и регрессии.

- LightGBM: быстрая и эффективная библиотека градиентного бустинга, используемая для классификации и регрессии.

- CatBoost: библиотека с открытым исходным кодом для градиентного бустинга, оптимизированная для категориальных данных.

- H2O.ai: коммерческая платформа машинного обучения, предоставляющая широкий спектр алгоритмов и инструментов для создания и развертывания моделей машинного обучения.

- Azure Machine Learning: облачная платформа машинного обучения от Microsoft, предоставляющая управляемую среду для разработки, обучения и развертывания моделей машинного обучения.

Эти инструменты машинного обучения предоставляют широкий спектр функциональных возможностей, включая:

- Предварительная обработка и очистка данных

- Выбор признаков и уменьшение размерности

- Обучение и оценка моделей машинного обучения

- Оптимизация гиперпараметров

- Развертывание и мониторинг моделей

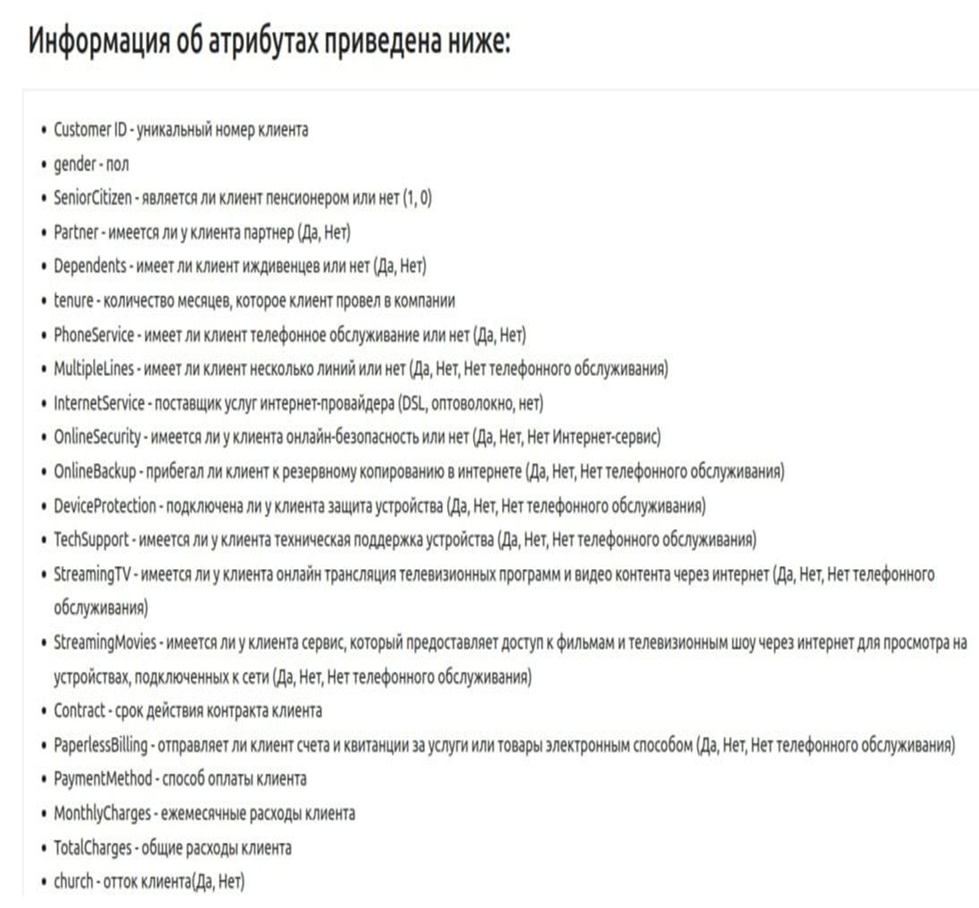
- Визуализация и отчетность

Инструменты машинного обучения значительно упрощают и ускоряют процесс разработки и развертывания моделей машинного обучения, позволяя специалистам по данным и разработчикам создавать более точные и эффективные модели.

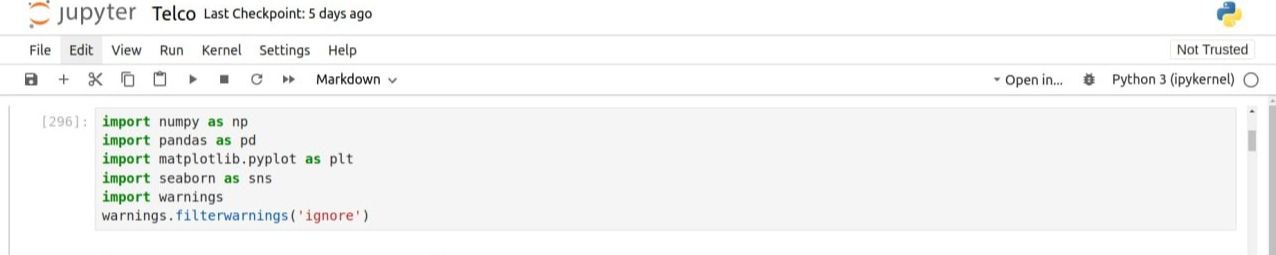
**Глава 2. Проведение анализа данных компании в Jupiter notebook**

**2.1 Исследовательский анализ данных в Jupiter notebook. Часть 1**

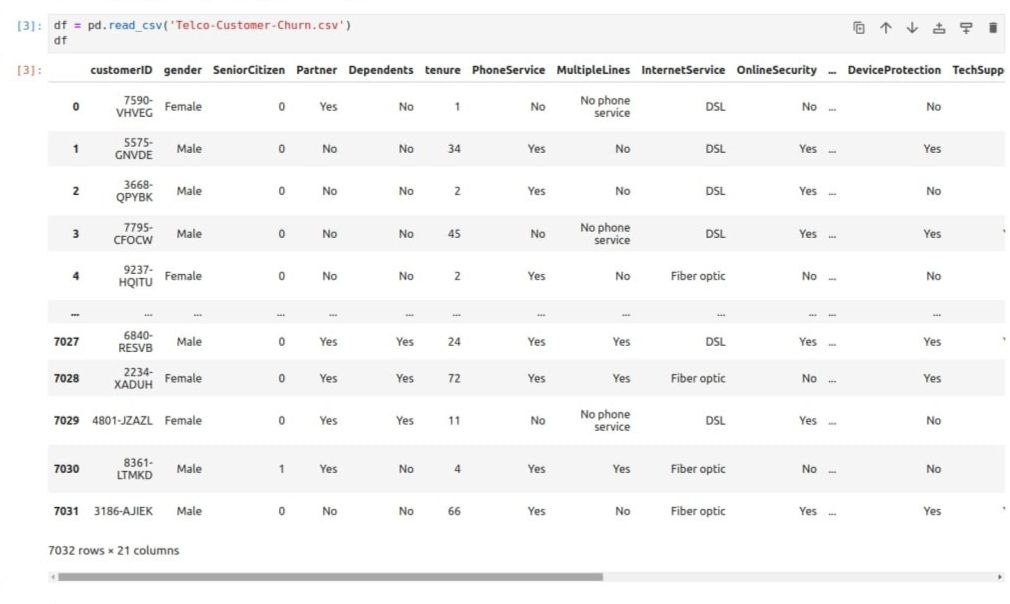
Опишем обозначения из файла WA\_Fn-UseC\_-Telco-Customer-Churn.csv.

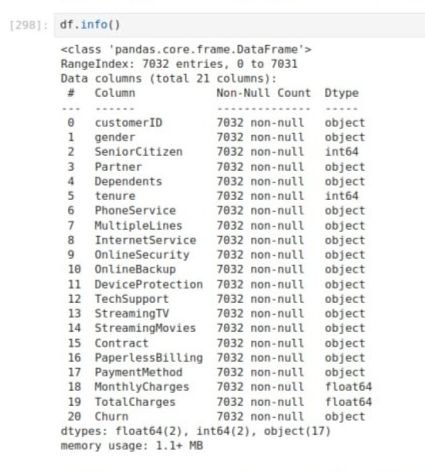


Импортируем необходимые библиотеки.



Импортируем исследуемый датасет WA\_Fn-UseC\_-Telco-Customer-Churn.csv, в котором представлена информация о различных переменных (описание выше) компании.

Проверим типы данных и есть ли отсутствующие данные в датасете.



Видно, что пропусков и отсутствующих данных не имеется.

Выведем статистические метрики для числовых колонок и посмотрим на закономерность.



Видно, что большинство колонок являются категориальными (нечисловыми), поэтому для них будем создавать dummy-переменные.

Проведём тщательный анализ. Убедимся ещё раз, что в исследуемых данных нет отсутствующих значений.

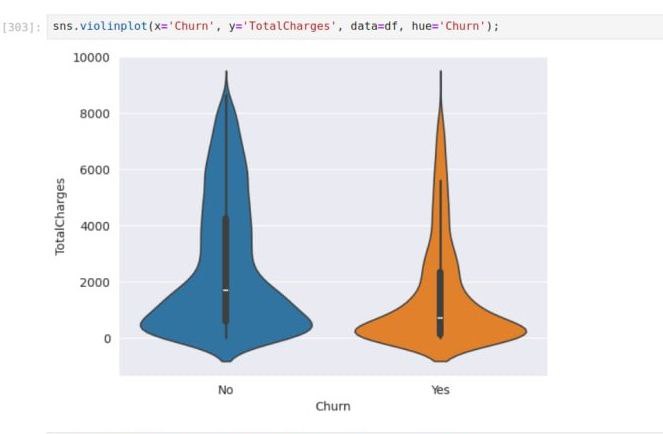


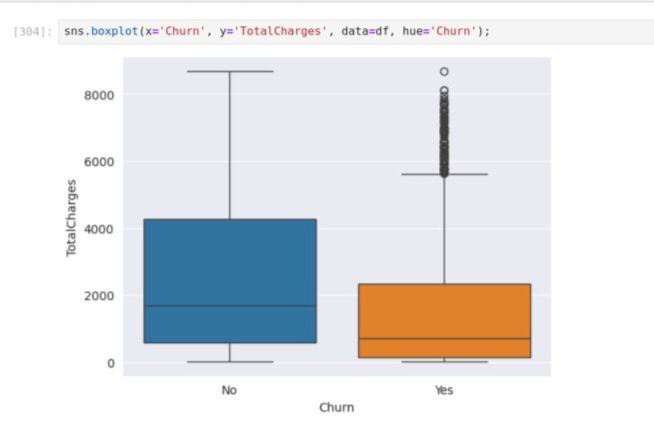
Мы получили все 0 на всех колонках, что означает отсутствующих значений нет. Теперь перейдём к визуализации данных. Построим график CountPlot для проверки сбалансированности значений колонок с классами (Churn).



Заметно сразу, что классы немного не сбалансированы. Клиентов, которые уходят в отток в 2.8 раз меньше, чем те, кто остаются в компании. Но данная разница между классами не является большой.

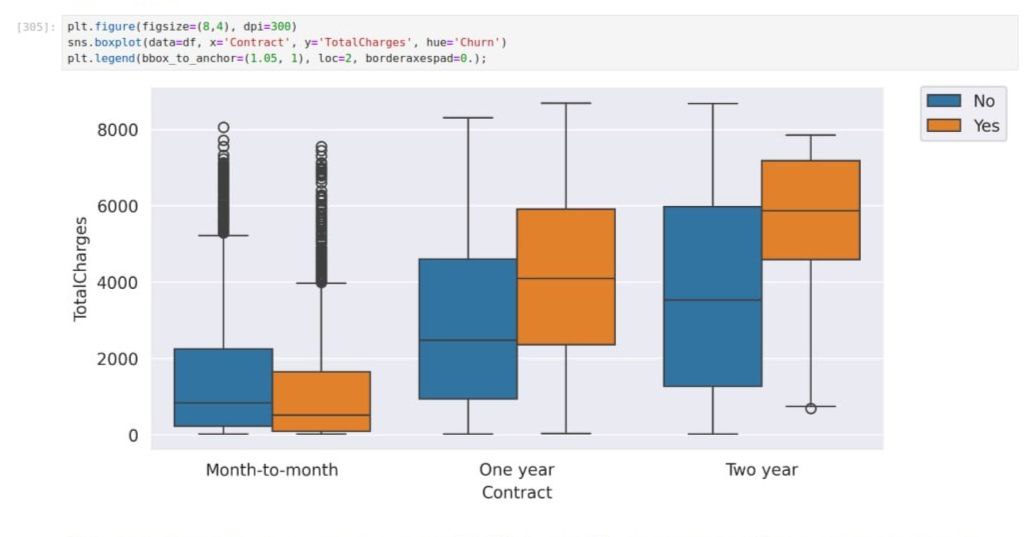
Теперь исследуем распределение колонки TotalCharges по различным категориям Churn с помощью графиков Box Plot и Violin Plot.





Здесь также видно, что есть перекос данных и несбалансированность значений. Наблюдаются определённое количество выбросов у тех, кто уходят в отток. К ним вернёмся позже.

Построим график boxplot с распределением колонки TotalCharges для различных типов контрактов.



Если мы посмотрим на тип контракта Month-to-month (контракты на определённое количество месяцев, то есть меньше года), то значение Churn в них примерно одинаковое. Почти схожая картина и для контрактов One year contract, а вот для контрактов Two year видно, что меньше клиентов уходят в отток.

Теперь подумаем, что могла бы сделать компания, чтобы меньше людей уходило в отток.

По вышеприведённым графикам видно, что чем больше значение TotalCharges (общие расходы) на любые виды контрактов, тем больше людей уходят в отток, и либо находят выгодные условия у конкурентов или, к примеру, не получают качественный сервис обслуживания в компании. Данный факт следует взять на вооружение.

Стоит задуматься также над тем, что почему при больших расходах большая часть клиентов не уходят в отток.

Построим график barplot с корреляцией числовыми признаками с целевой переменной (Churn). Для категориальных признаков проведём конвертацию их в dummy-переменные, так как корреляция вычисляется только для числовых колонок.



Отметим, что мы выделим в переменной Churn только атрибут Churn\_Yes (уход в отток) и применим её при построении графика barplot.

На основе этого теперь построим график barplot.



В итоге получим следующий рисунок:

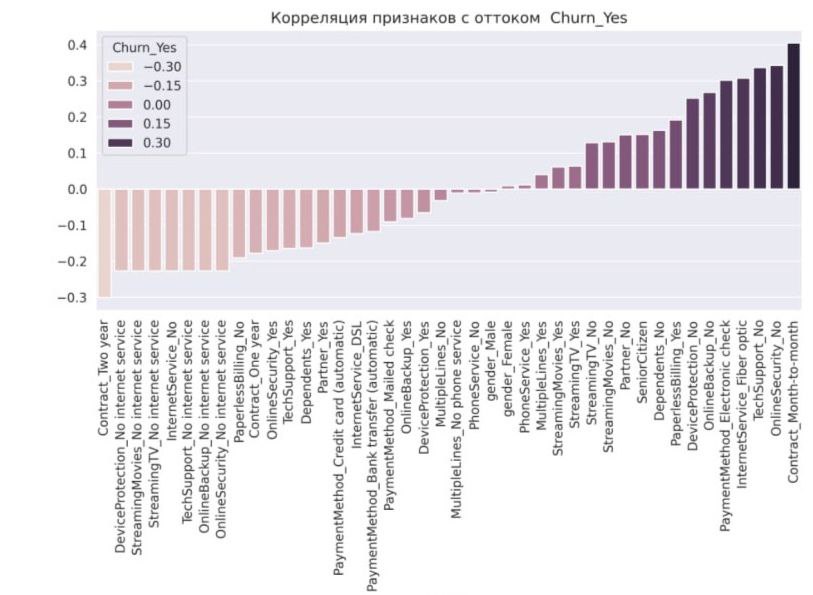


График корреляции с атрибутом Churn\_Yes показывает, что для краткосрочных контрактов (Contract\_Month-to-month) с большими расходами характерен высокий отток клиентов. Тут прямая зависимость: с возрастанием значения Contract\_Month-to-month, возрастает в количественном соотношении Churn\_Yes.

Мы выяснили, что длительность контракта по времени оказывает немаловажное значение на отток клиентов и соответственно будем анализировать влияние всех видов контрактов на отток клиентов из компании.

Проведём анализ оттока и когорт. Разобьём всех абонентов(клиентов) на отдельные сегменты по длительности обслуживания в компании. Это нам поможет понять, как меняется поведение клиентов в зависимости от срока обслуживания.

Посмотрим какие у нас есть виды контрактов.



Посмотрим сколько месяцев клиент являлся или является абонентом в компании и также посмотрим какое количество клиентов по каждому месяцу.



Видна следующая закономерность по признаку tenure (количество месяцев, которое клиент провёл в компании): чем больше длительность контракта по времени, тем меньше клиентов заключают договор с компанией. Хотя эта зависимость не строгая, наблюдаются определённые цикличности, но всё же следует заострить внимание на краткосрочных контрактах.

Построим гистограмму с распределением значений колонки tenure.



Видно, что очень много клиентов имеют значение tenure от 1 до 2 месяцев. Второе пиковое значение находится в районе 72. На графике отчётливо заметны всплески между первым и вторым пиковыми значениями, что скорее означает продление годовых и двухлетних контрактов.

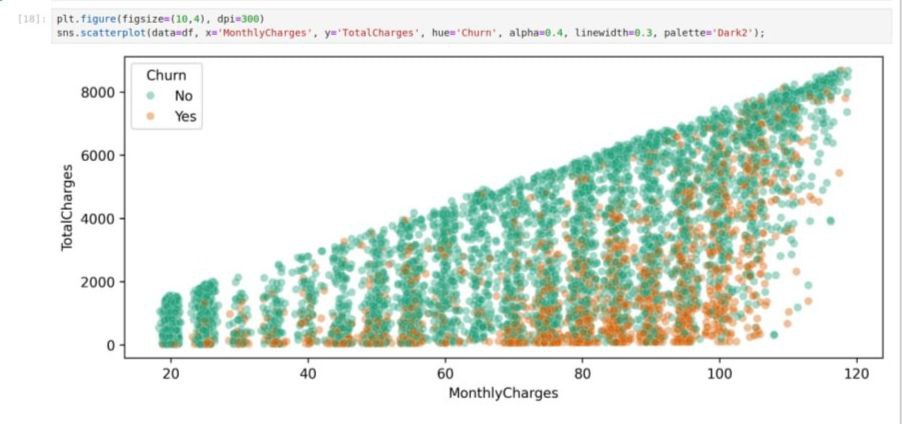
Теперь построим график displot по признаку tenure, колонке Contract, строке Churn и посмотрим на получившийся результат.



По графику displot видно, что на годовых и двухгодовых контрактах, уходящих в отток клиентов мало, поэтому основные усилия можно направить на контракты 'Month-to-month'. Причём для краткосрочных контрактов (меньше года) пик значения, наблюдается от 1 до 2 месяцев для клиентов, уходящих в отток. Но затем идёт пологое распределение. Следует учесть тот факт, что ранее мы наблюдали большой отток клиентов из-за больших расходов на долгосрочных контрактах. Возможно, отток на контрактах 'Month-to-month' обусловлен из-за высоких цен (стоимость тарифа) чем на долгосрочных контрактах. В таком случае кампании следовало бы проводить активную маркетинговую акцию, т.е. предлагать клиентам с контрактом 'Month-to-month' перейти на годовые или двухгодовые контракты, тем самым доход от этих абонентов снизится, но в долгосрочной перспективе кампания получит больший суммарный доход от этого абонента.

Таким образом, анализ графика displot поможет определить стратегию по снижению оттока клиентов и увеличению доходов компании. Важно учитывать особенности каждого типа контракта и предлагать клиентам наиболее выгодные и удобные условия для них.

Построим график scatterplot для колонок Total Charges и Monthly Charges, колонки раскрасим по колонке Churn.



Очередной раз подтверждается тот факт, что в отток уходят те, у кого большие расходы. В принципе, это логично, чем больше ежемесячные расходы, тем больше у клиентов возникает причин искать более дешёвые тарифы у конкурентов.

Да, и вообще, это очень дискуссионный вопрос: "Что для кампании важнее: либо абонент с ежемесячными высокими платежами (MonthlyCharges), либо абонент с высокими общими платежами (TotalCharges)". Эти два показателя связаны между собой через длительность по времени(tenure).

На графике scatterplot мы подтверждаем, что в отток уходят клиенты с высокими ежемесячными расходами, причём в основном у них небольшие значения (TotalCharges).

Это может быть связано с тем, что клиенты с большими ежемесячными расходами могут быть более ценными для компании в долгосрочной перспективе, если они остаются на услугах компании длительное время. Однако, клиенты с высокими общими платежами могут приносить компании больше прибыли в краткосрочной перспективе.

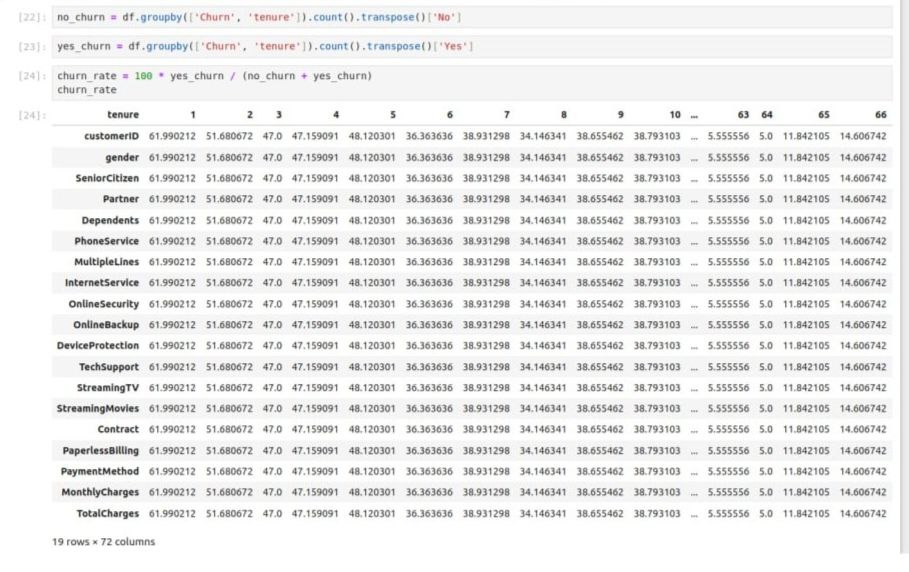
Таким образом, важно учитывать оба показателя при анализе оттока клиентов и разработке стратегии удержания.

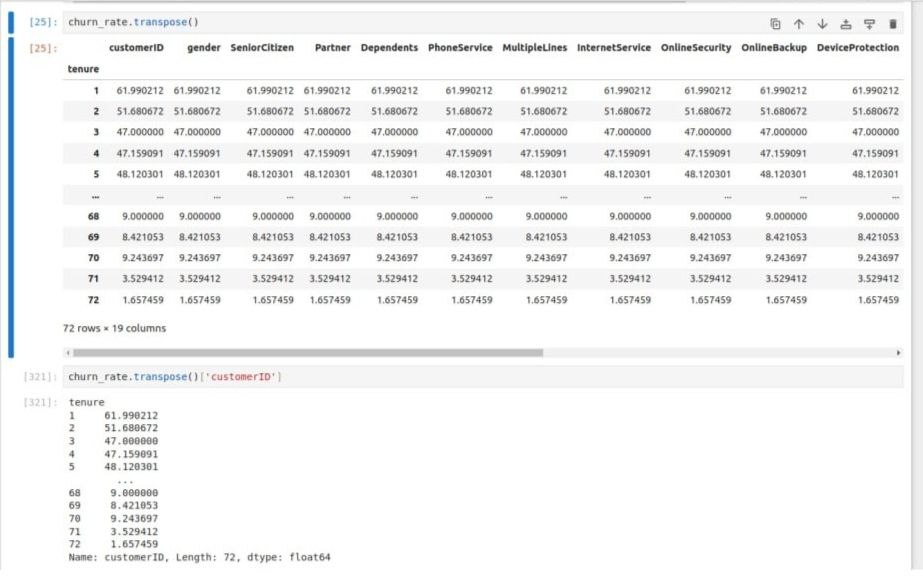
**2.2 Исследовательский анализ данных в Jupiter notebook. Часть 2**

Создадим когорты(сегменты) по колонке tenure. Для начала поместим различные значения tenure (1 месяц, 2 месяц, 3 месяца и так далее) в отдельные когорты(сегменты).

Для каждого уникального значения колонки tenure в качестве когорты, вычислим процент оттока (churn rate) – это количество тех людей, кто ушёл в отток в когорте, как процент от общего количества людей в данной когорте. Вычислим такой процент отдельно для каждой когорты. В итоге должно получиться когорты от 1 до 72 месяцев, и с увеличением количества месяцев уровень оттока должен снижаться. Это разумно, потому что чем дольше клиент пользуется услугами кампании, тем вероятнее то, что его/её всё устраивает, и он/она продолжит пользоваться этими услугами.

Чтобы выполнить разбиение отдельные когорты по каждому месяцу, проведём ряд операций и преобразований. Отдельно разобьём кто уходит в отток(yes\_churn) и тех, кто остаётся в компании(no\_churn). И по этому разбиению уже определим, процент ушедших по каждому месяцу.





Построим график, показывающий отток клиентов за каждый месяц.



Видно, что на графике пологое распределение. Чем больше значение tenure, тем меньше процент оттока в когорте. Отчётливо также видно, до 12 месяцев отток в процентном соотношении большой. После 12 месяцев идёт постепенное снижение оттока. Очередной раз убеждаемся, что упор компании нужно сделать на контракты до года.

Оптимизируем процесс исследования. Будем двигаться в следующем ключе. На основе колонки tenure создадим колонку с названием Tenure Cohort, в которой будут следующие 4 категории:

- 0-12 месяцев

- 12-24 месяца

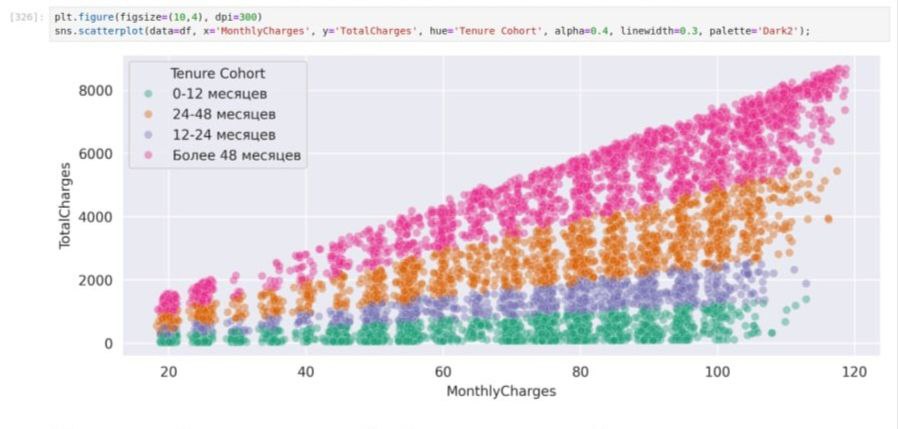
- 24-48 месяцев

- Более 48 месяцев

На основе этого создадим функцию для преобразования данных, которая будет иметь следующий вид:

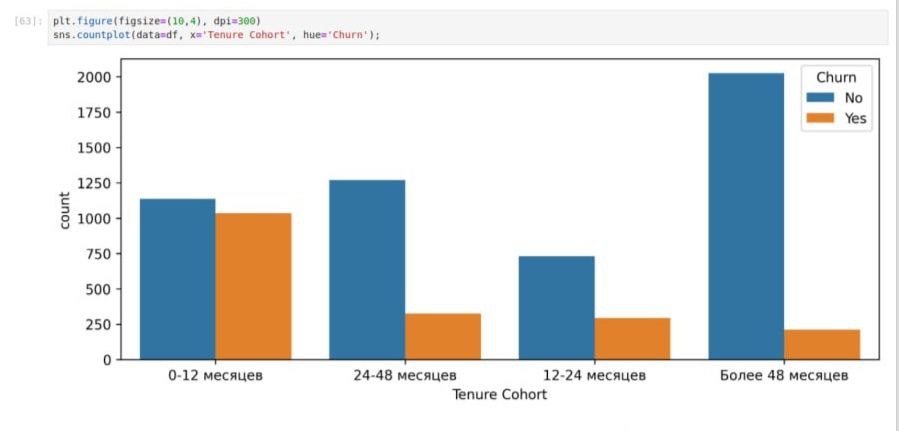


Построим график scatterplot для Total Charges и Monthly Charges, раскрашивая график разными цветами по колонке Tenure Cohort, которую создали ранее.



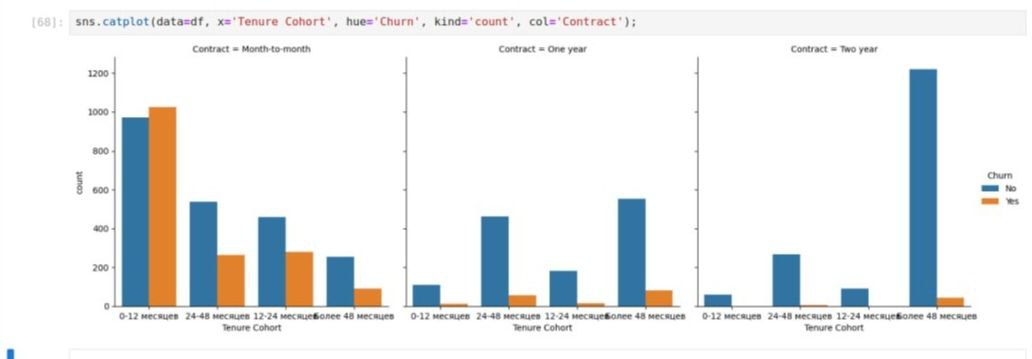
На рисунке всё выглядит логично. Чем больше когорта, тем большее значение TotalCharges она принимает. Сам рисунок является более читаемым и легко воспринимаемым для анализа.

Построим график countplot с количеством ушедших и не ушедших в отток людей в каждой когорте.



Отчётливо видно, что клиенты с длительностью обслуживания более 48 месяцев менее подвержены к оттоку, а те, кто пользуются услугами кампании менее 12 месяцев, то вероятность оттока заметно возрастает.

Создадим набор графиков CountPlot с количеством клиентов по различным когортам Tenure Cohort. Для каждого типа контракта создадим отдельный график, и раскрасим эти графики разными цветами по колонке Churn.



Вышеприведённый график, как и другие, также подтверждает тот факт, что клиенты на однолетнем и двухлетнем контрактах меньше уходят в отток вне зависимости от длительности пользования услугами. Об этом свидетельствует то, что процент оттока на контрактах продолжительностью в один и два года остается ниже, чем на контрактах другой продолжительности, как на начальных этапах пользования услугами, так и на более поздних. Вероятно, клиенты, заключившие контракт на год или два, более лояльны и удовлетворены услугами провайдера.

**Глава 3. Применение моделей машинного обучения для прогнозирования оттока клиентов**

**3.1 Выбор модели машинного обучения**

Теперь мы подошли к основной части нашей работы – создание предиктивной модели машинного обучения. Возникает невольный вопрос: “Какую модель следует выбирать?”

Сделаем небольшое напоминание, что при выборе модели машинного обучения необходимо учитывать следующие факторы:

1. Цель задачи: определить, какую задачу мы хотим решить (классификация, регрессия, кластеризация и т. д.).

2. Объем и характер данных: учесть размер выборки, количество признаков, их тип (категориальные или числовые) и распределение.

3. Время обучения и предсказания: учесть требования к скорости работы модели.

4. Интерпретируемость: если важно понимать, как модель принимает решения, то лучше выбрать модели с хорошей интерпретируемостью.

5. Регуляризация: если имеется много признаков, которые могут быть нерелевантными, рассмотреть модели с возможностью регуляризации.

6. Разреженность данных: если есть разреженные данные (многие нулевые значения), выбрать модели, которые умеют с ними хорошо работать.

7. Проверка модели: использовать кросс-валидацию и метрики качества, чтобы выбрать наилучшую модель.

8. Наличие предобученных моделей: если у нас ограниченные ресурсы, то лучшем рассмотреть использование предобученных моделей для нашей задачи.

Исходя из этих факторов, можно выбрать наиболее подходящую модель машинного обучения для нашей конкретной задачи. Важно также помнить, что иногда эффективно использовать не одну модель, а ансамбль моделей для повышения качества предсказаний.

В нашем случае данные размечены, что означает классифицированы и соотнесены к определённому классу. Здесь уместно использовать обучение с учителем. Выбор наш падёт на модели на основе деревьев, логистическая регрессия, метод опорных векторов (SVM), метод K-ближайших соседей (KNN).

Данные модели машинного обучения имеют несколько преимуществ:

1. Интерпретируемость: легко понимать и интерпретировать, поскольку они представляют собой простую логическую структуру. Это делает их полезными для принятия решений и объяснения прогнозов.

2. Универсальность: могут быть использованы как для задач классификации, так и для задач регрессии. Они могут быть применены к данным с различными типами признаков, включая категориальные признаки.

3. Работа с нелинейными зависимостями: могут обрабатывать сложные нелинейные зависимости в данных без необходимости их линеаризации.

4. Устойчивость к выбросам: хорошо устойчивы к выбросам и шуму в данных.

Также модели на основе деревьев, логистическая регрессия, метод опорных векторов (SVM), метод K-ближайших соседей (KNN) могут быть применены в различных областях, включая:

1. Классификация: часто используются для классификации объектов на основе их признаков. Например, они могут быть применены в медицине для диагностики заболеваний, в финансах для прогнозирования рисков и мошенничества, или в маркетинге для выявления целевой аудитории.

2. Регрессия: могут использоваться для прогнозирования непрерывных значений. Например, они могут быть применены в финансовой аналитике для прогнозирования цен на акции или в промышленности для прогнозирования времени до отказа оборудования.

3. Кластеризация: могут использоваться для кластеризации данных и выявления групп похожих объектов.

4. Ранжирование: могут быть использованы для ранжирования объектов в зависимости от их значимости или релевантности.

Использование этих моделей даст целый спектр ответов, которые мы хотим узнать.

Исследуем шесть моделей на основе классификации: одно дерево решений, случайный лес, адаптивный бустинг, градиентный бустинг, логистическая регрессия, метод опорных векторов (SVM), метод K-ближайших соседей (KNN), а далее сравним результаты и выберем из них наилучшую с меньшей ошибкой предсказания (в нашем случае метрика recall).

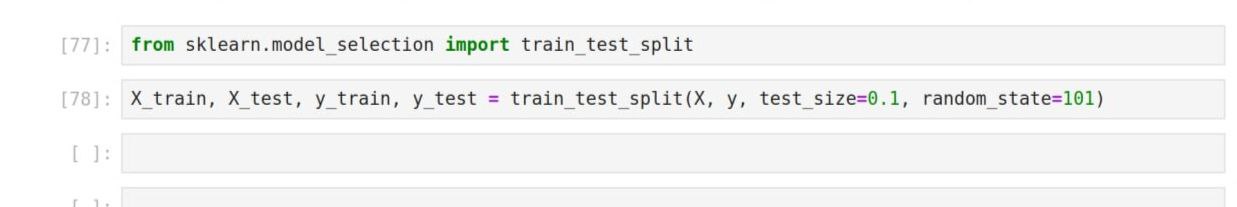
**3.2 Модель Decision Tree**

Модель Decision Tree (одно дерево решений) - метод машинного обучения, который использует структуру дерева для принятия решений. Каждый узел дерева представляет собой условие, которое определяет, какие данные нужно анализировать, а каждое ребро соответствует возможному результату этого условия. Путем прохождения по дереву от корня к листьям можно получить прогноз для конкретного наблюдения. Decision Tree является простым и интерпретируемым методом, который может быть эффективно использован для задач классификации и регрессии.

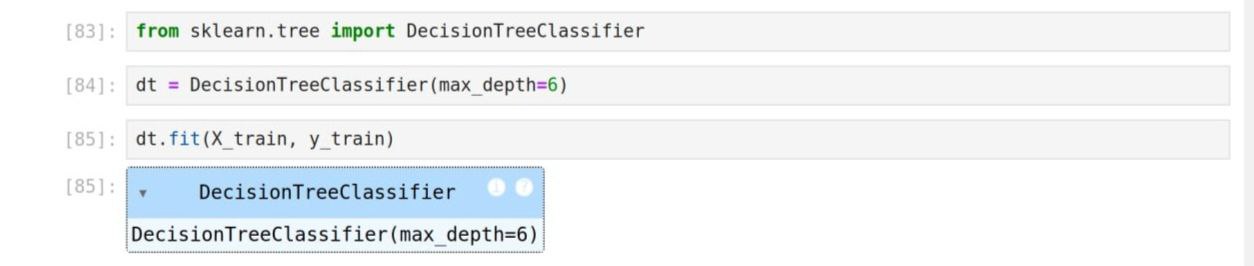
Разделим все данные на признаки Х и целевую переменную У. Создадим dummy-переменные, где это необходимо, а также посмотрим, есть ли какие-то признаки, которые можно было бы удалить.



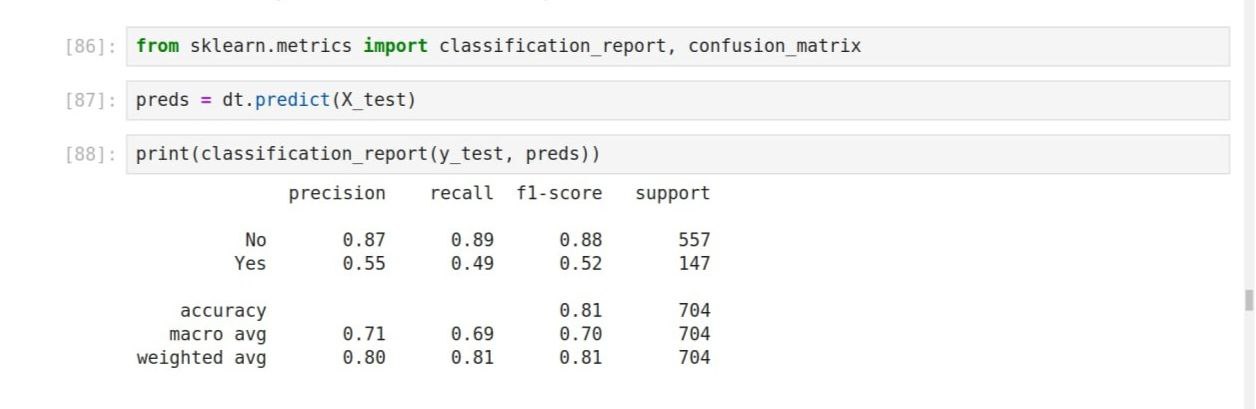
Выполним разбиение данных на обучающий и тестовый наборы (train\_test\_split), откладывая в сторону 10% данных для тестирования. В решениях мы используем random\_state=101.



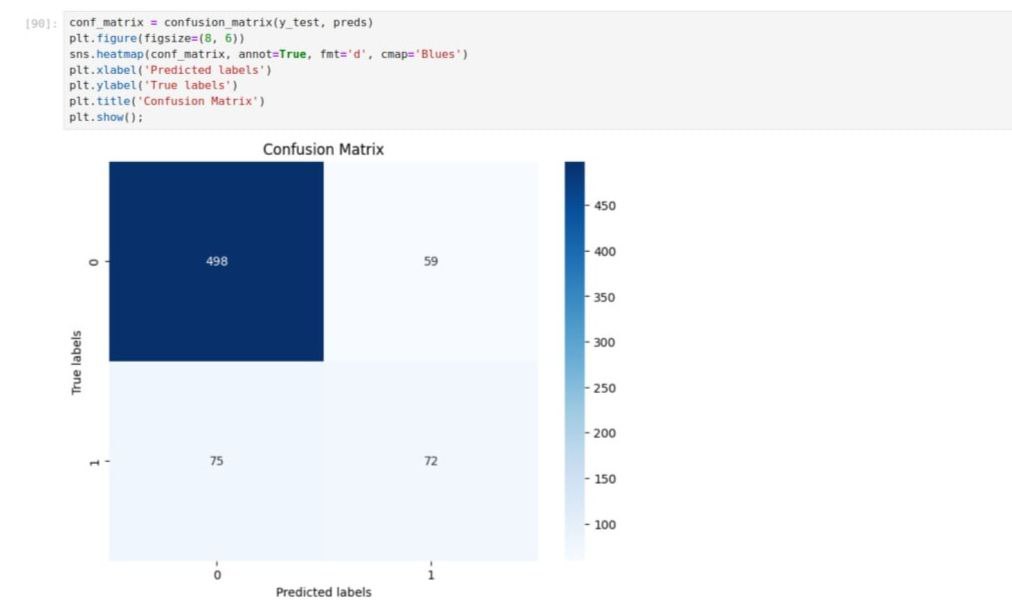
Построим модель Decision Tree, оценим её работу и выполним её обучение.



Оценим метрики работы модели дерева решений, выведим отчёт "classification report" и нарисуем график с матрицей ошибок (confusion matrix).

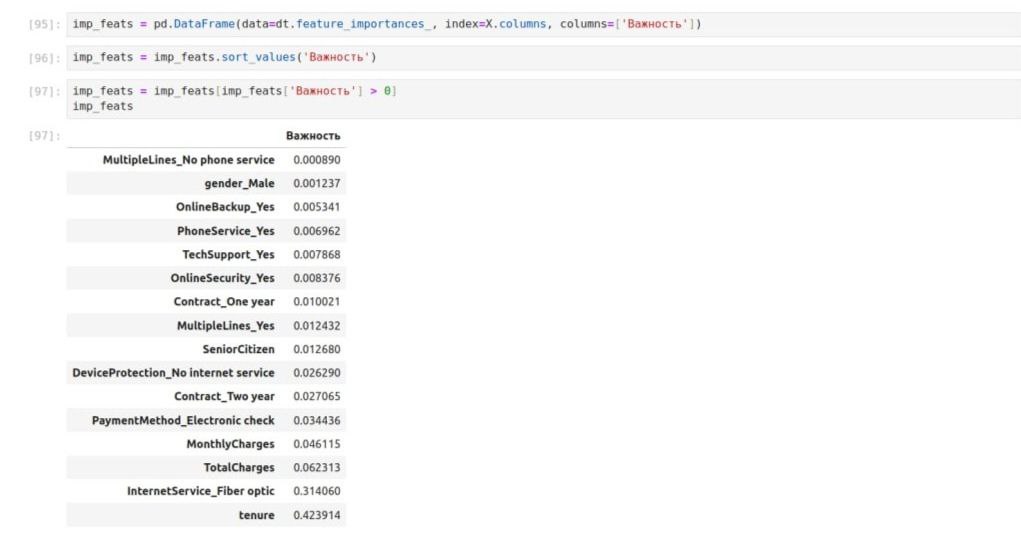


По данным видно, что модель лучше работает для класса Churn['No'] чем Churn['Yes']. Это логично и разумно, так как в классе Churn['No'] содержится больше данных.

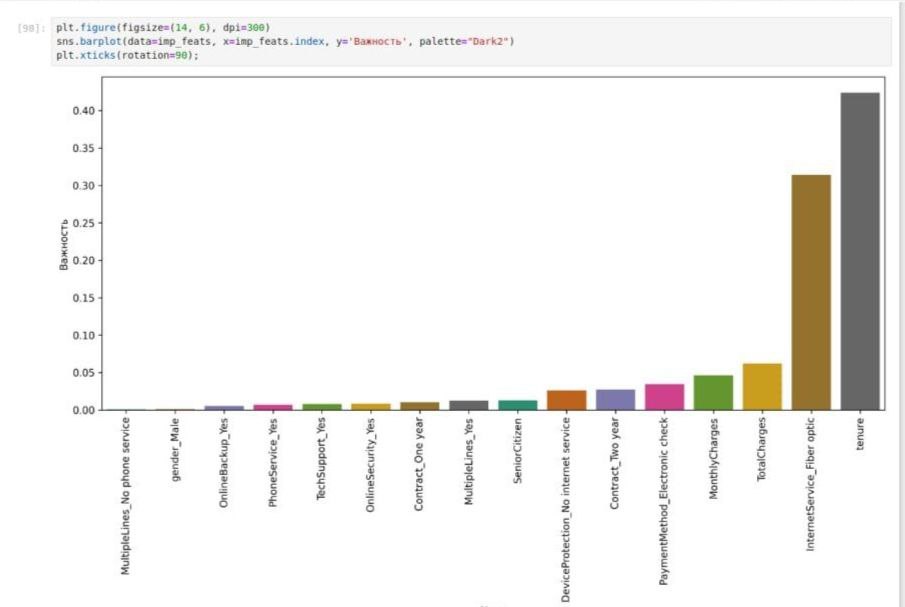


Мы видим, что неправильно классифицированы 75 абонентов, когда наша модель предсказала, что оттока не будет, но отток случился и именно это число мы хотели бы минимизировать, так как здесь мы упускаем тех, кто уходит в отток. Кроме этого, есть ещё 59 людей, для которых модель предсказала, что они уйдут в отток, но на самом деле они не ушли в отток.

Оценим, какие признаки играют наибольшую роль в принятии решений моделью.



Визуализируем полученные данные.



Ранее мы видели, что признак tenure хорошо коррелирует с целевой переменной. Однако, есть ещё один интересный признак Internetservice\_Fiber optic, который ранее не встречали нигде в исследовании. В этом случае нужно выяснять отдельно, почему клиенты, которые нуждаются в интернете по оптоволокну, могут уходить в отток. Возможно, у кампании выше цены, чем у конкурентов, именно по этой услуге.

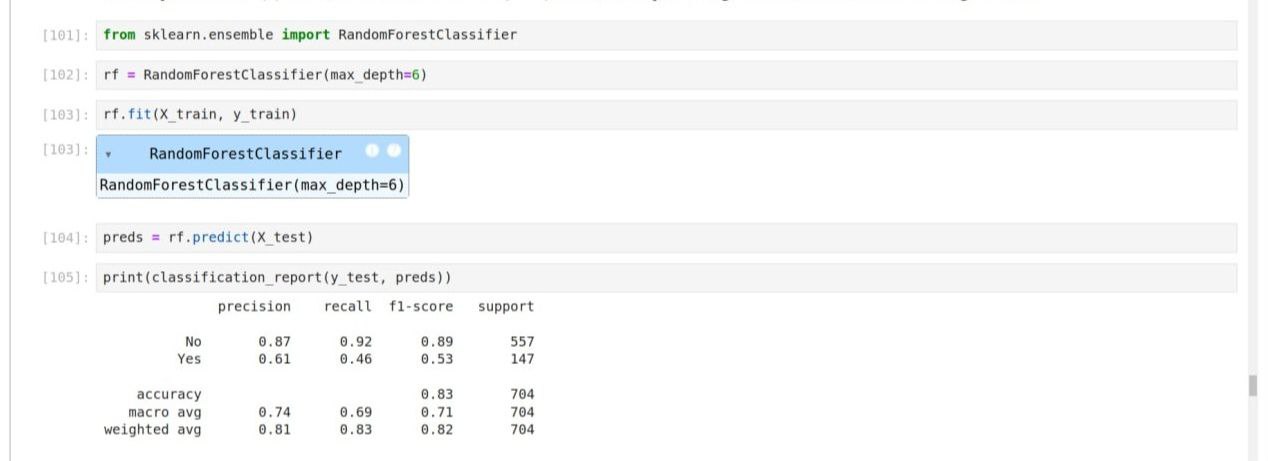
К примеру, для проверки данной гипотезы можно провести анализ цен на услуги интернет - провайдера и его конкурентов, сравнить их тарифные планы и узнать, предлагают ли конкуренты более привлекательные условия для клиентов, нуждающихся в интернете по оптоволокну.

Также стоит провести опрос клиентов, уходящих в отток, чтобы узнать, почему они решают сменить интернет - провайдера. Вопросы опроса могут касаться стоимости услуг, качества обслуживания, скорости интернета и других факторов, которые могут повлиять на принятие решения об уходе.

Исходя из результатов анализа цен, тарифных планов и опроса клиентов, можно будет сделать вывод о том, является ли цена услуг интернета по оптоволокну одной из причин ухода клиентов в отток от. Если это подтвердится, компания сможет принять меры для улучшения своего конкурентного положения на рынке и удержания клиентов.

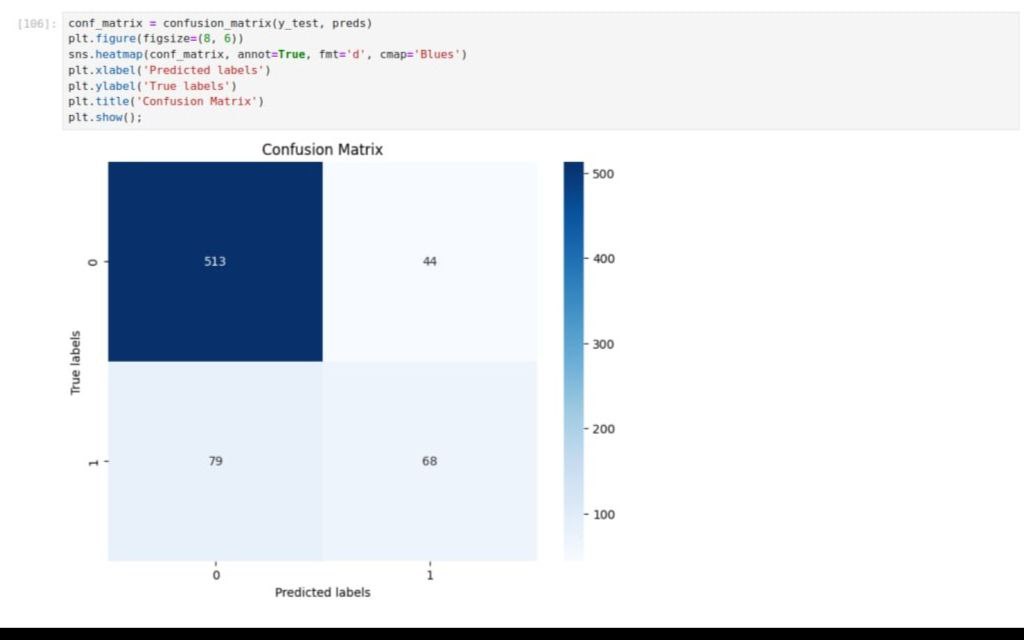
**3.3 Модель Random Forest**

Модель Random Forest (случайный лес) – это алгоритм машинного обучения, который использует комбинацию нескольких деревьев принятия решений для решения задач классификации или регрессии. Он работает путем создания леса случайных деревьев во время обучения и совершения предсказаний на основе голосования или усреднения результатов всех деревьев. Random Forest обладает высокой точностью, устойчив к переобучению и способен обрабатывать как категориальные, так и количественные данные.



Видно, что эта модель работает немного хуже, чем модель Decision Tree (важно счесть, какая метрика учитывается и сравнивается). Причина в том, что дерево решений получилось слишком глубоким, и из-за этого модель может получится переобученной (overfitting), т.е. хорошо работать на обученных данных и плохо работать на тестовых данных. Поэтому в определённых ситуациях полезно ограничивать глубину дерева (max\_depth).

Построим матрицу ошибок и посмотрим на результат.



В матрице ошибок (Confusion Matrix) мы видим интересную картину. Эта модель имеет более высокие метрики accuracy и precision с поправкой на ограничение(max\_depth=6). Однако, если опираться на нашу задачу, то мы хотим минимизировать количество тех абонентов, которые по факту ушли в отток, но наша модель думает, что они не уйдут в отток. В данной матрице ошибок это значение равно 79, хотя раньше это значение было равно 75 для модели одного дерева. И за это число отвечает метрика recall, и эту метрику нам необходимо минимизировать. Вывод: модель случайного леса показывает худшее значение метрики recall, чем модель дерево решений.

Поэтому, если основной задачей является минимизация количества ложно отрицательных результатов (когда модель предсказывает, что клиент останется, но он на самом деле уходит), то лучше использовать модель дерева решений с глубиной равной 6. В данном случае accuracy и precision могут быть не такими высокими, как у модели случайного леса, но recall будет лучше.

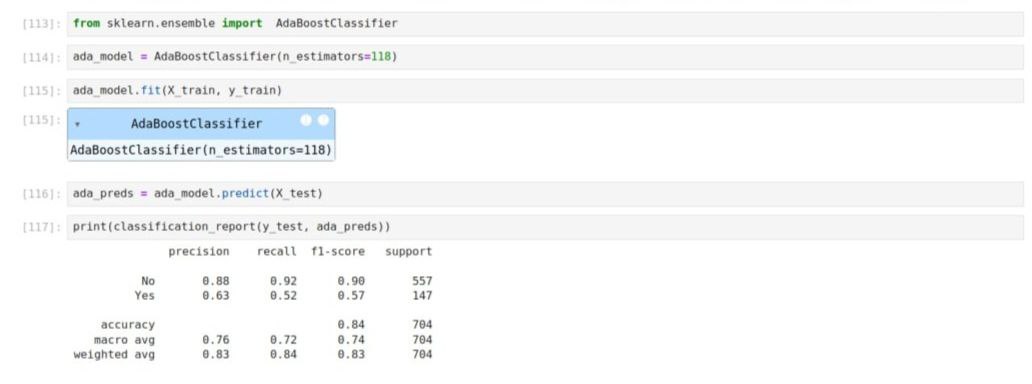
Но делать поспешных выводов пока не стоит. Рассмотрим ещё две модели, но уже с применением гиперпараметров.

**3.4 Модель Boosted Trees**

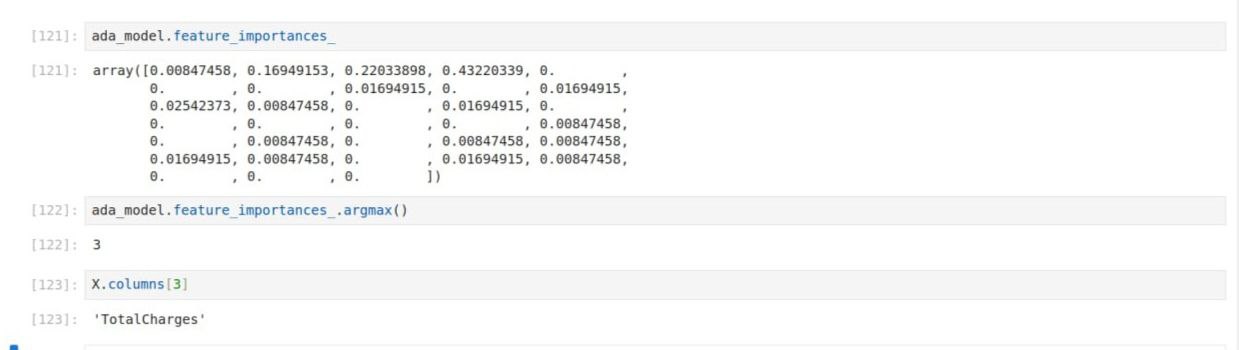
Модель Boosted Trees (расширяемые деревья) – это ансамблевый метод машинного обучения, который объединяет несколько слабых моделей (деревьев решений) в одну сильную модель.

Мы построим поочерёдно модели AdaBoost и Gradient Boosting.

Создадим модель AdaBoost с применением гиперпараметров, оценим её работу и выполним обучение.

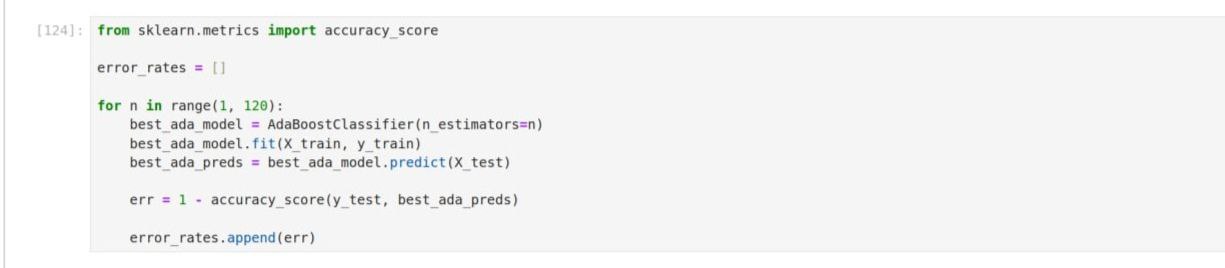


Оценим важность признаков. Посмотрим какие из признаков оказывают наибольшее влияние на целевую переменную в данной модели.



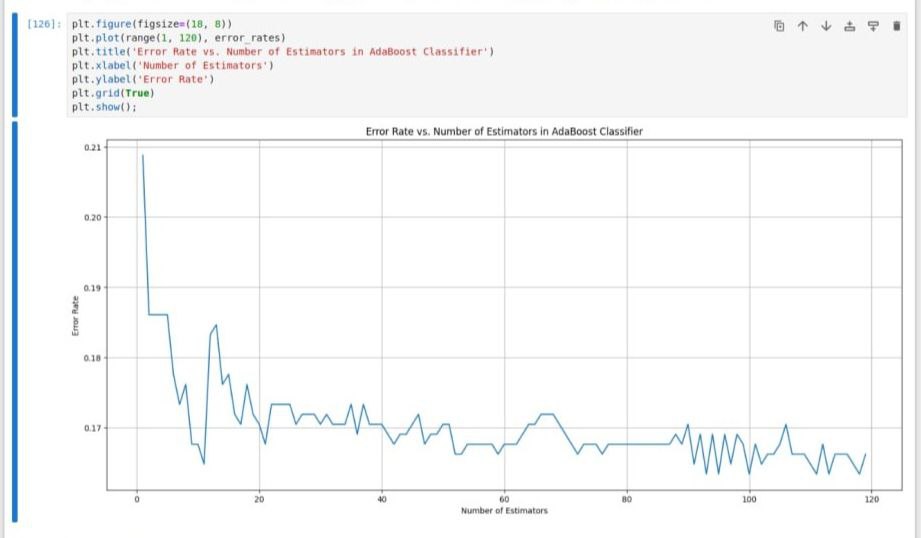
Получили, что в модели AdaBoost наиболее важное значение имеет признак TotalCharges. И это справедливо, так как показывает сколько всего расходов понёс клиент в кампании, и также данный признак достаточно хорошо коррелирует с целевой переменной Churn. Поэтому можно сделать вывод, что суммарные расходы клиента являются важным показателем для прогнозирования оттока клиентов.

Оптимизируем модель.



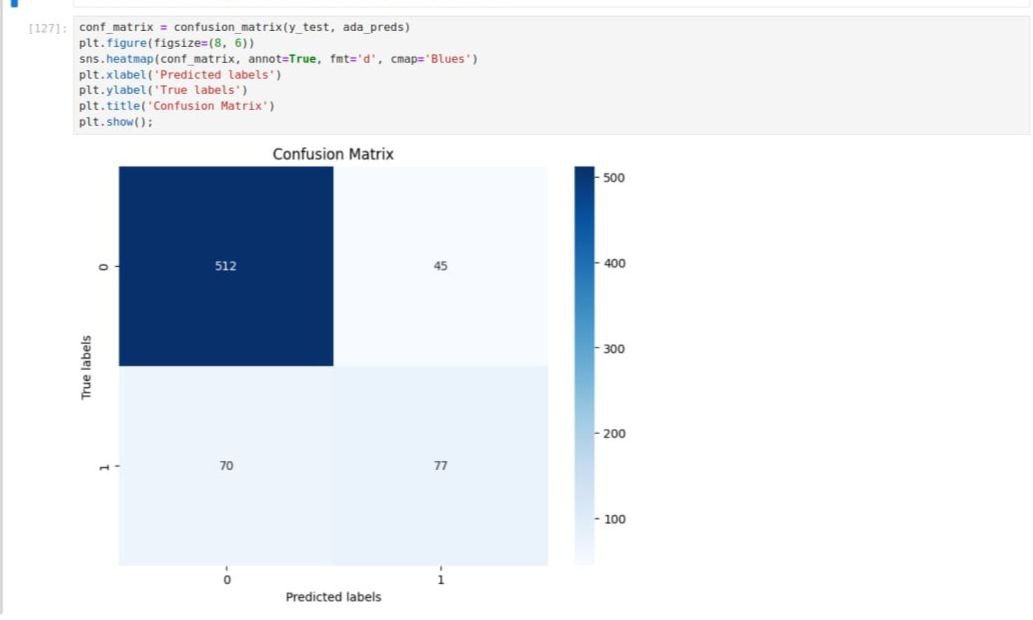
При увеличении количества деревьев ошибка модели должна уменьшатся, так как увеличение количества деревьев в модели может помочь уменьшить ошибку на тренировочном наборе данных, что может привести к уменьшению ошибки на тестовом наборе данных. Однако, если количество деревьев становится слишком большим, модель может начать переобучаться и ошибаться на новых данных. Поэтому всегда важно подбирать оптимальное количество деревьев в модели.

Визуализируем наши данные и посмотрим на результат.



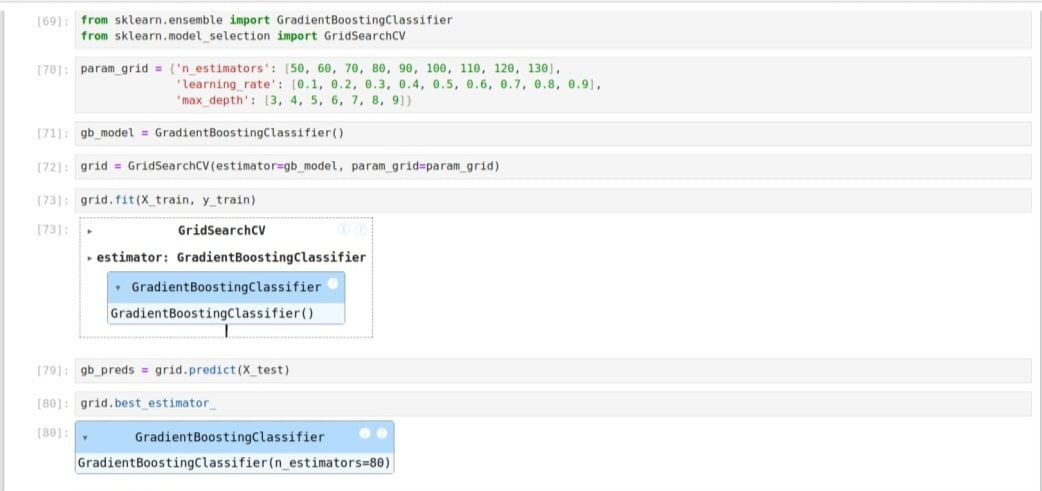
По графику минимальное значение ошибки находится где-то в районе 118. Это значение мы и указали при обучении модели выше. Это значение ошибки является оптимальным для данной модели и набора данных, так как дальнейшее уменьшение ошибки может привести к переобучению модели и ухудшению ее предсказательной способности на новых данных. Поэтому важно найти баланс между снижением ошибки и предотвращением переобучения при обучении модели.

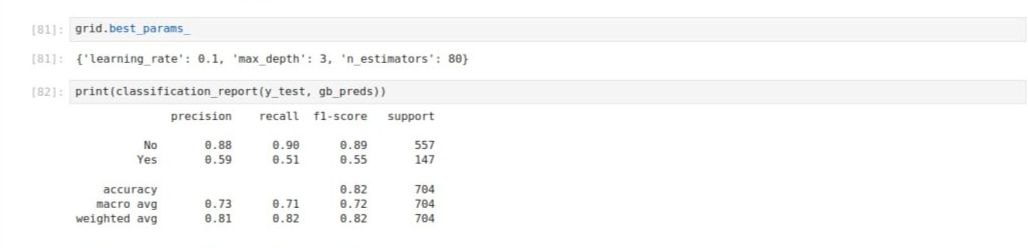
Построим матрицу ошибок для данной модели.

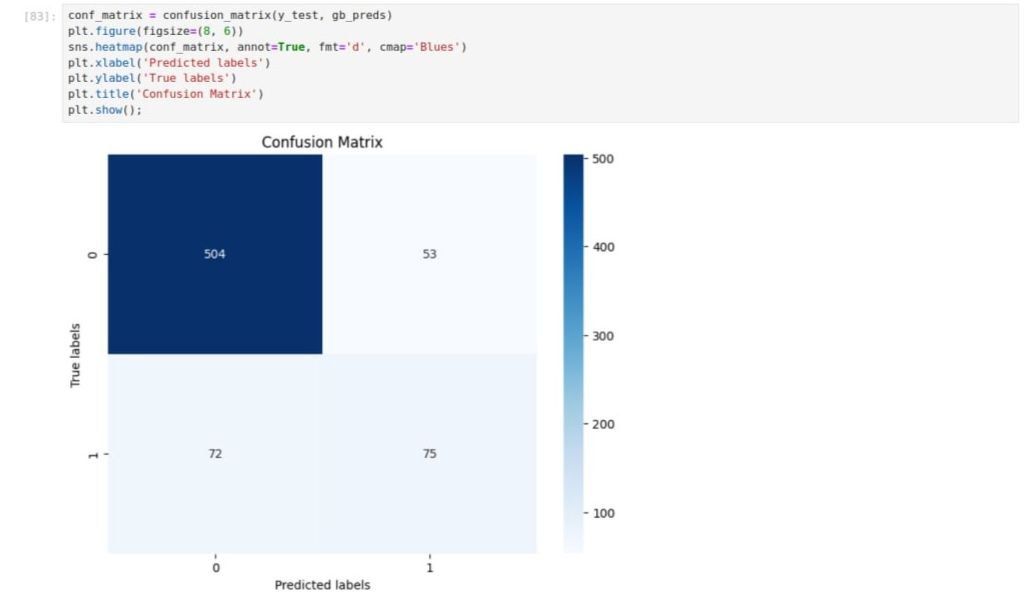


Мы уменьшили значение метрики recall. И это значение равно 70. Это наилучшее значение из всех, что мы получали ранее на других моделях. Пока данная модель имеет лучшую предсказательную способность.

Создадим модель Gradient Boosting с применением гиперпараметров, оценим её работу и выполним обучение.







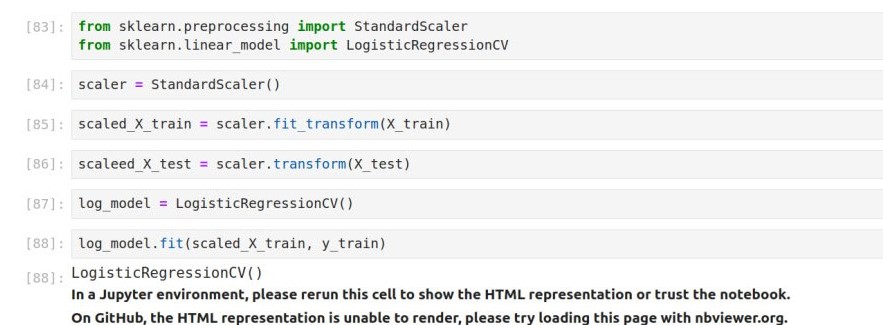
Следует отметить, что обучение данной модели проходит очень долго по сравнению с другими тремя моделями. Скорее это связанно, что поиск оптимальных значений происходит по сетке (GridSearchCV), и в модели мы указали много входных данных в качестве гиперпараметров.

В итоге модель Gradient Boosting отработала чуть хуже, чем предыдущая модель.

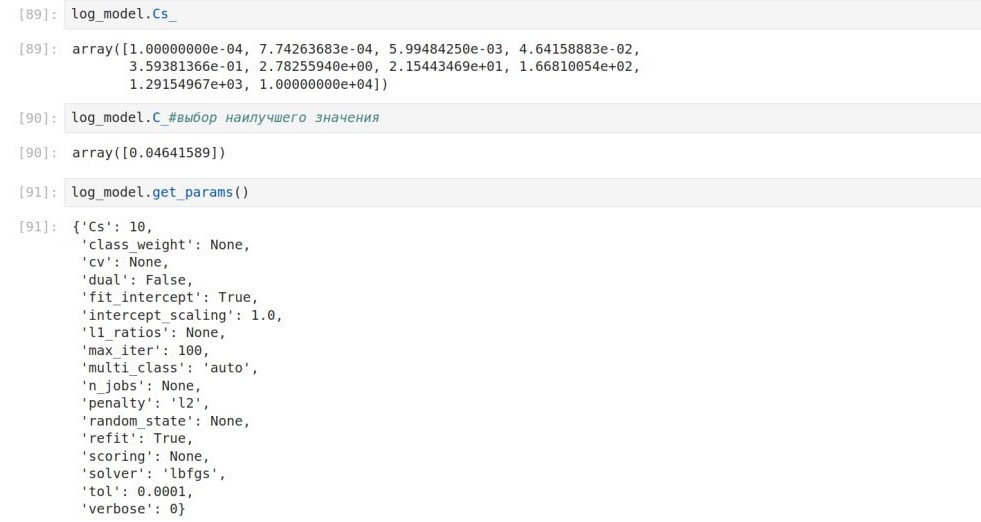
**3.5 Модель логистической регрессии**

Модель логистической регрессии – это статистический метод, который используется для предсказания вероятности того, что наблюдение принадлежит к определенной категории, исходя из одного или нескольких предикторов.

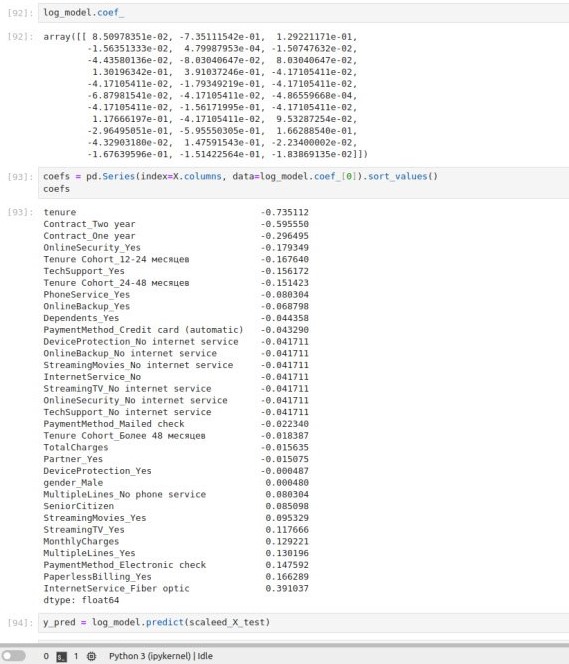
Построим модель логистической регрессии и обучим её.



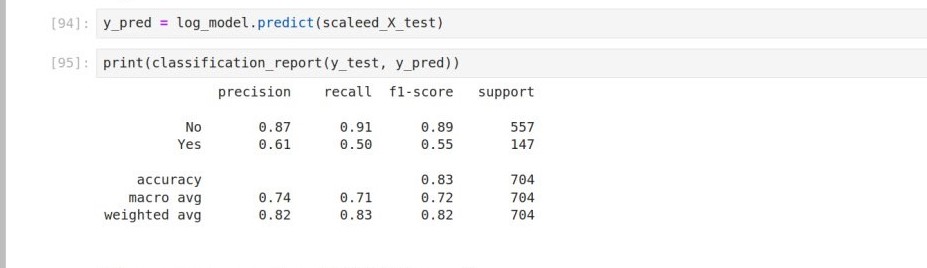
Посмотрим на оптимальные значения данной модели.

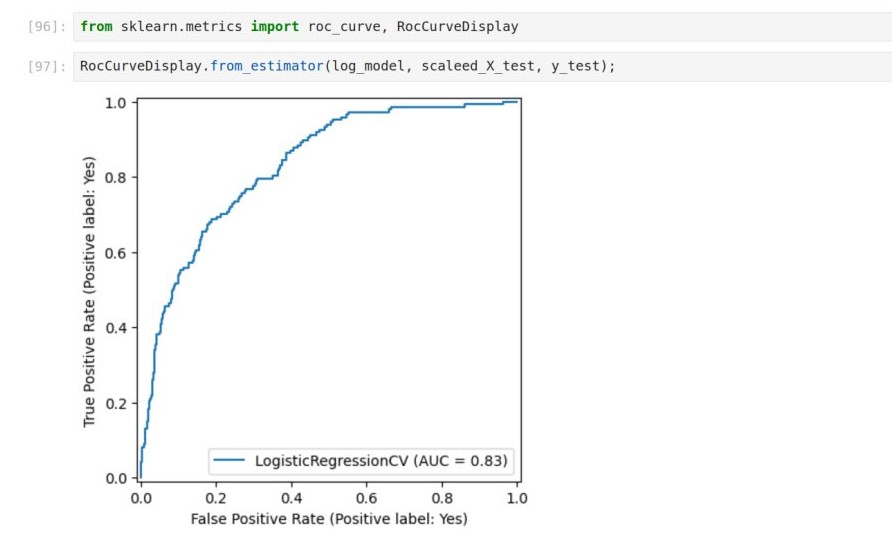


Посмотрим на коэффициенты модели.



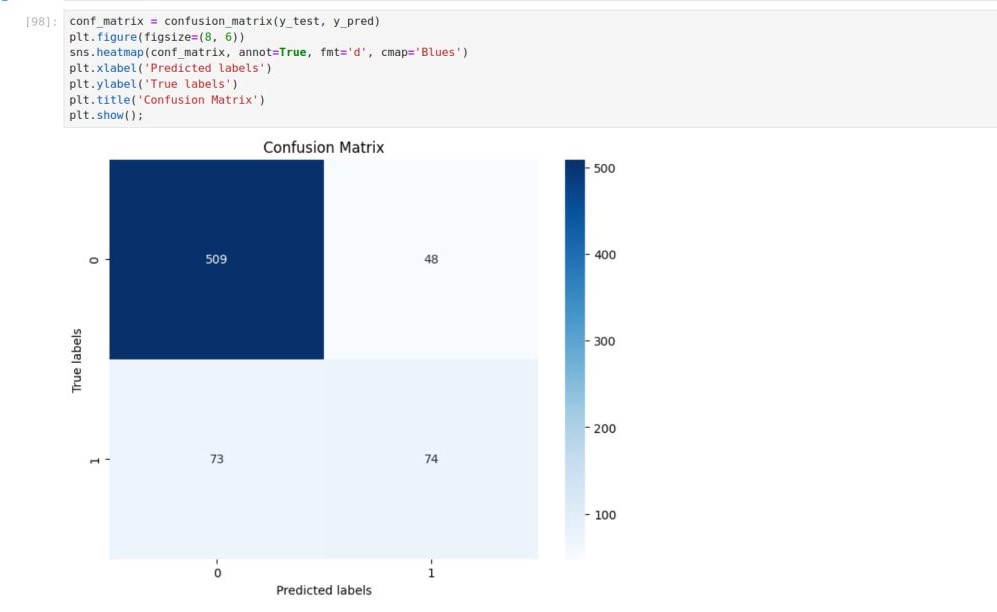
Выполним предсказание модели и построим отчёт классификации.



Построим график кривой логистической регрессии (ROC CURVE)

LogisticRegressionCV (AUC=0.83) в нашем случае означает, что была применена модель логистической регрессии с кросс-валидацией, и значение площади под кривой ROC (AUC) для данной модели составляет 0.83. AUC является метрикой, которая измеряет качество классификационной модели, где значение ближе к 1 указывает на более точную модель. Таким образом, значение AUC=0.83 указывает на хорошее качество модели логистической регрессии.

Построим матрицу ошибок и посмотрим на значение метрик.

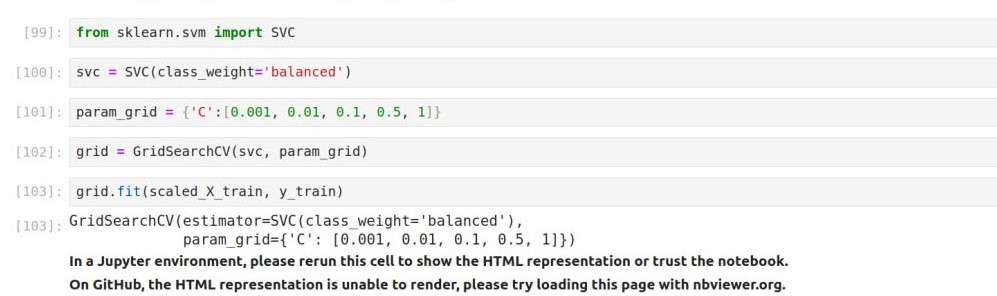


Видно, что логистическая регрессия немного хуже отрабатывает, чем модель AdaBoost. Метрика recall равна 73. Ранее наилучшее значение мы получали 70.

**3.6 Метод опорных векторов (SVM)**

Метод опорных векторов (SVM) — это алгоритм машинного обучения, используемый для классификации и регрессии. Он особенно эффективен для задач с высокой размерностью и малым количеством обучающих данных.

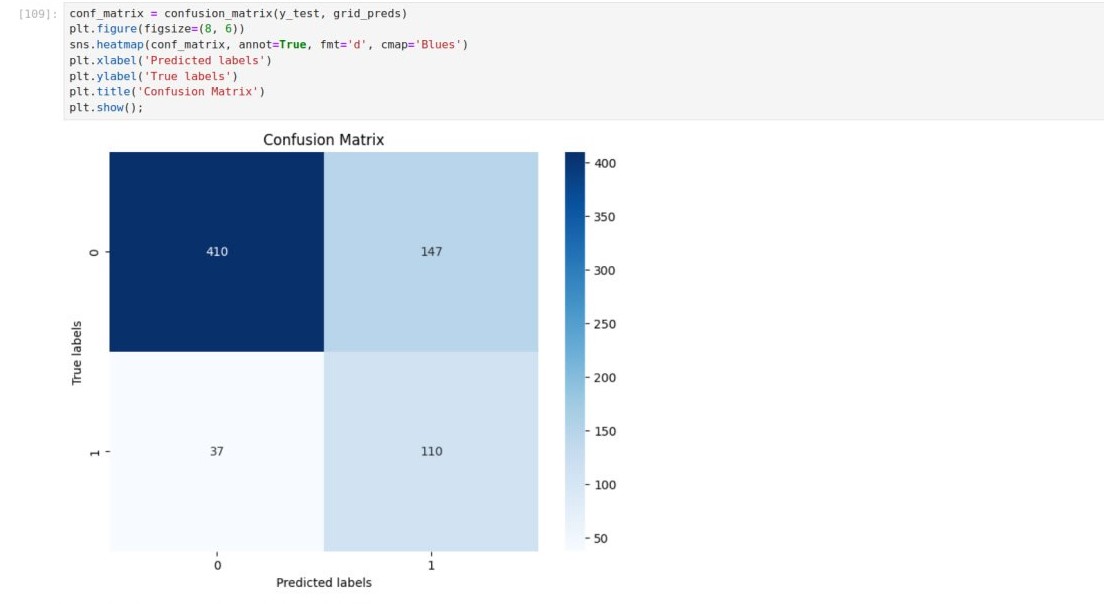
Выполним обучение модели.



Определим лучшие параметры данной модели.



Построим матрицу ошибок и сделаем вывод.

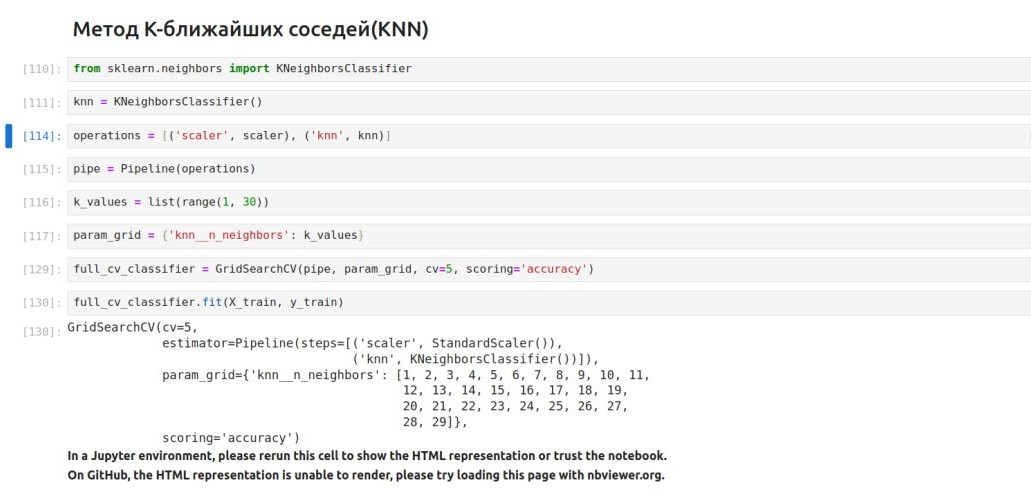


Мы получили наилучшее значение recall равной 37. Уменьшили ошибку предсказания для этой метрики почти в 2 раза, по сравнению с предыдущими моделями. Предсказание данной модели в целях решения нашей задачи является пока лучшей.

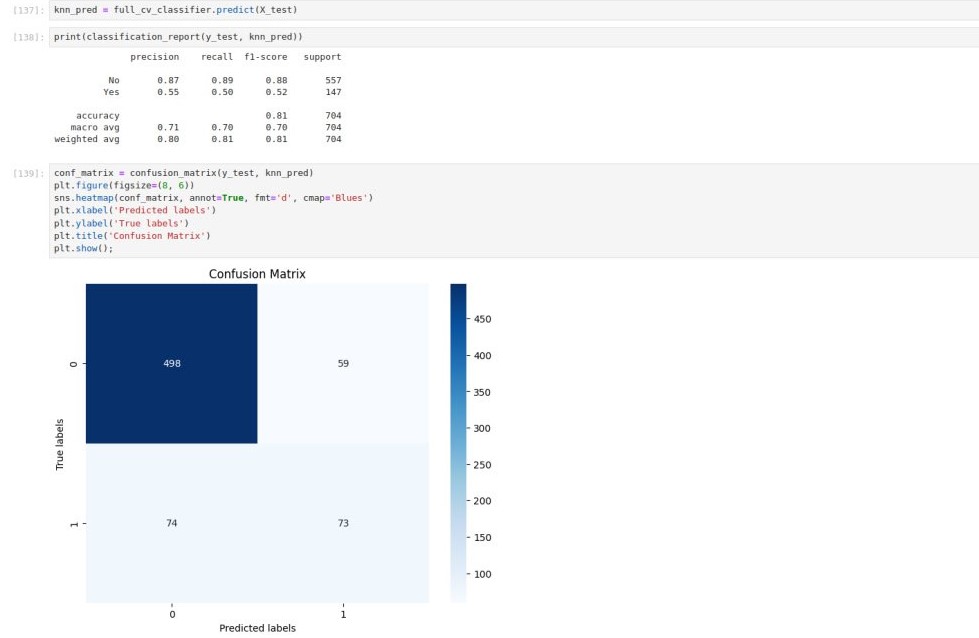
**3.7 Метод K-ближайших соседей (KNN)**

Метод K-ближайших соседей (KNN) — это алгоритм машинного обучения, который используется для классификации или регрессии. Он основан на идее, что точки данных, которые расположены близко друг к другу в пространстве признаков, скорее всего, относятся к одному и тому же классу.

Построим данную модель для наших данных.



Выполним обучение модели и построим матрицу ошибок.



Значение recall (равное 74) получилось достаточно большим. Это хуже, чем у многих ранее рассмотренных моделей.

**3.8 Общий вывод**

Если сделаем вывод, то можно сказать, что модель метод опорных векторов (SVM) с использованием гиперпараметров работает лучше остальных моделей. Значение метрики recall у неё наименьшее чем у других моделей. И соответственно, предсказательная способность у неё намного лучше. Эту модель можно использовать в нашей задаче, ведь у неё и ошибка предсказания меньше. И наша модель готова. Её можно использовать в дальнейшем (при добавлении и появлении новых данных в датасете, модель нужно будет корректировать).

Было выяснено, что наибольшее влияние на отток клиентов влияет показатель TotalCharges (общие расходы), и больше оттока наблюдается на краткосрочных контрактах. В связи с этим, компании рекомендуется пересмотреть условия краткосрочных контрактов (меньше года), чтобы удержать абонентов у себя. Исходя из этого, следует разработать некоторые мероприятия, которые удовлетворяли бы клиентов на контрактах меньше года.

**Заключение**

В данном исследовании было проведено прогнозирование оттока клиентов в телекоммуникационной компании с использованием алгоритмов машинного обучения. Целью исследования было выявить факторы, влияющие на отток клиентов, и разработать рекомендации по их удержанию.

Для достижения этой цели был проведен анализ данных клиентов, включающий информацию о демографических характеристиках, услугах, использовании ресурсов и удовлетворенности клиентов. Затем были применены различные алгоритмы машинного обучения, такие как модели на основе деревьев, логистическая регрессия, метод опорных векторов (SVM), метод K-ближайших соседей (KNN) для построения моделей прогнозирования оттока.

Результаты исследования показали, что на отток клиентов влияют несколько ключевых факторов. Одним из них является тип услуги, которую клиент использует. Например, клиенты, использующие только интернет-услуги, имеют более высокую вероятность оттока. Кроме того, длительность пользования услугами компании и уровень удовлетворенности клиентов также оказывают значительное влияние на отток. Наибольший отток наблюдался на краткосрочных контрактах.

На основе этих результатов были составлены рекомендации по удержанию клиентов. Во-первых, компания должна обратить особое внимание на клиентов, использующих только интернет-услуги, и предложить им дополнительные услуги, чтобы повысить их удовлетворенность и уменьшить вероятность оттока. Во-вторых, компания должна проводить регулярные опросы клиентов для выявления и устранения возможных проблем и недовольств. Кроме того, компания может предлагать персонализированные скидки и бонусы клиентам, которые пользуются услугами компании в течение длительного времени, чтобы стимулировать их оставаться.

Таким образом, применение алгоритмов машинного обучения для прогнозирования оттока клиентов и составление рекомендаций по их удержанию может значительно помочь телекоммуникационным компаниям, таким как, в улучшении своей стратегии удержания клиентов и повышении своей конкурентоспособности на рынке.