МИНОБРНАУКИ РОССИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ»

(СПбГЭУ)

Факультет экономики и финансов

Кафедра прикладной математики и экономико-математических методов.

КУРСОВАЯ РАБОТА

по дисциплине

«Методы и средства интеллектуального анализа данных»

Тема «Прогнозирование оттока клиентов сети фитнес-центров с использованием методов машинного обучения»

Направление 38.03.01 «Экономика»

Профиль: Математические методы и анализ данных в экономике

Обучающийся Намозова Томирис

Группа Э-2110 Подпись

Проверила Леора Светлана Николаевна

Должность к. ф.-м. н., доцент кафедры ПМ иЭММ

Оценка Дата:

Подпись:

Санкт-Петербург

2024

Содержание

[ВВЕДЕНИЕ 4](#_Toc185957714)

[1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ. ЗАДАЧА БИНАРНОЙ КЛАССИФИКАЦИИ И ОБЗОР МЕТОДОВ КЛАССИФИКАЦИИ В МАШИННОМ ОБУЧЕНИИ 5](#_Toc185957715)

[1.1 Логистическая регрессия 5](#_Toc185957716)

[1.2 Метод опорных векторов (SVM) 6](#_Toc185957717)

[1.3 Метод ближайших соседей (KNN) 7](#_Toc185957718)

[1.4 Случайный лес (Random Forest) 8](#_Toc185957719)

[2. ПРАКТИЧЕСКАЯ ЧАСТЬ. РЕАЛИЗАЦИЯ МОДЕЛЕЙ НА ДАННЫХ 10](#_Toc185957720)

[2.1. Знакомство и предобработка данных 10](#_Toc185957721)

[2.2. Реализация моделей 17](#_Toc185957722)

[2.3 Оценка важности признаков 27](#_Toc185957723)

[2.4. Вывод результатов и их анализ 28](#_Toc185957724)

[3. ОБЩИЕ ВЫВОДЫ И РЕКОМЕНДАЦИИ 30](#_Toc185957725)

[3.1 Важность частоты посещений 30](#_Toc185957726)

[3.2 Продолжительность времени с момента первого обращения 30](#_Toc185957727)

[3.3 Срок до окончания действующего абонемента 31](#_Toc185957728)

[3.4 Возраст клиентов 31](#_Toc185957729)

[3.5 Использование дополнительного дохода 31](#_Toc185957730)

[ЗАКЛЮЧЕНИЕ 33](#_Toc185957731)

[Список использованной литературы 34](#_Toc185957732)

[Приложение А – Проверка на пропуски и дубликаты. 35](#_Toc185957734)

[Приложение Б – Убрали верхний регистр столбцов 35](#_Toc185957735)

[Приложение В – Предобработка данных 36](#_Toc185957736)

[Приложение Г - Разделение данных на признаки (матрица Х) и целевую переменную (у). Формируем обучающую и тестовую выборки. 36](#_Toc185957737)

[Приложение Д - Определяем функции для обучения и оценки моделей 36](#_Toc185957738)

[Приложение Е – Логистическая регрессия 37](#_Toc185957739)

[Приложение Ж – Метод опорных векторов 37](#_Toc185957740)

[Приложение З – Метод ближайших соседей 38](#_Toc185957741)

[Приложение И – Случайный лес 38](#_Toc185957742)

[Приложение К - Общий график ROC AUC и матрицы ошибок для каждой модели 39](#_Toc185957743)

# **ВВЕДЕНИЕ**

В последние годы наблюдается устойчивая тенденция увеличения конкуренции в сфере фитнес-услуг, что ведет к необходимости более глубокого понимания потребностей клиентов и их поведения. Одна из ключевых задач для успешного функционирования и развития фитнес-центров заключается в прогнозировании оттока клиентов. Удержание клиентов является важным аспектом бизнеса, поскольку затраты на привлечение новых клиентов зачастую значительно выше, чем на удержание существующих. Поэтому важно выявлять факторы, способствующие уходу клиентов, и разрабатывать эффективные стратегии предотвращения оттока.

Современные методы машинного обучения предоставляют мощные инструменты для анализа данных и прогнозирования различных событий, включая отток клиентов. Эти методы позволяют обрабатывать большие объемы информации, выявлять скрытые закономерности и строить предсказательные модели на основе исторических данных. Автоматизация процесса анализа клиентского поведения с помощью машинного обучения может значительно повысить эффективность принятых решений и, как следствие, улучшить финансовые показатели фитнес-центров.

В данной курсовой работе будет проведен анализ оттока клиентов фитнес-центров с использованием методов машинного обучения. Рассмотрены будут различные алгоритмы и методы оценки качества моделей, включая ROC-кривую и матрицы ошибок. Цель исследования заключается в разработке прогностической модели, которая позволит фитнес-центрам не только предвидеть потери клиентов, но и разрабатывать меры по их удержанию.

Таким образом, настоящая работа направлена на анализ ключевых факторов, способствующих оттоку клиентов, и формирование практических рекомендаций по их предотвращению с использованием передовых методов машинного обучения.

# **1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ. ЗАДАЧА БИНАРНОЙ КЛАССИФИКАЦИИ И ОБЗОР МЕТОДОВ КЛАССИФИКАЦИИ В МАШИННОМ ОБУЧЕНИИ**

Бинарная классификация представляет собой одну из важнейших задач в области машинного обучения, которая заключается в разделении объектов на две категории. Такая задача актуальна в сферах, где необходимо принимать решения на основе неполных или неструктурированных данных. В контексте прогнозирования оттока клиентов фитнес-центров бинарная классификация позволяет выделить клиентов, которые рискуют прекратить посещение центра, и тех, кто будет продолжать пользоваться услугами.

Решение задачи бинарной классификации обычно состоит из нескольких этапов:

- сбор данных: необходимо собрать информацию о клиентах;

- предобработка данных: этот этап включает очистку данных, обработку пропусков, кодирование категориальных признаков и нормализацию числовых значений;

- обучение модели: использование алгоритмов машинного обучения для обучения на тренировочных данных;

- оценка модели: проверка качества модели с использованием тестовых данных и метрик, таких как точность, полнота, F1-мера и ROC-AUC;

- выбор наилучшей модели и оценка важности признаков.

В области машинного обучения существует множество методов, применимых для решения задачи бинарной классификации. Рассмотрим наиболее распространенные из них:

## **1.1 Логистическая регрессия**

Логистическая регрессия представляет собой один из самых распространённых методов бинарной классификации, который используется для моделирования вероятности наступления события. Она строит логит-модель, что позволяет преобразовать линейную комбинацию входных переменных в вероятностное значение, расположенное в интервале от 0 до 1. Это достигается с помощью логистической функции (сигмоиды), которая применяется к линейному предсказанию.

Математическая основа

Модель логистической регрессии можно описать следующим образом:

(1)

где — вероятность принадлежности к положительному классу, — признаки, а — коэффиценты, которые мы стремимся оценить.

Преимущества:

- интерпретируемость: логистическая регрессия позволяет легко оценить влияние отдельных признаков на вероятность целевого события;

- скорость обучения: метод быстро обучается даже на больших объемах данных;

- отсутствие предполагаемого распределения: в отличии от классической линейной регрессии, логистическая регрессия не требует, чтобы переменные были нормально распределены.

Недостатки:

- линейные зависимости: логистическая регрессия наилучшим образом работает, когда существует линейная взаимосвязь между признаками и логитом вероятности;

- проблемы с многократной коллинеарностью: высокая корреляция между признаками может негативно сказаться на стабильности коэффициентов модели.

## **1.2 Метод опорных векторов (SVM)**

Метод опорных векторов (Support Vector Machine, SVM) — это мощный алгоритм для решения задач классификации и регрессии, который находит оптимальную гиперплоскость для разделения классов в пространстве признаков. Главное отличие SVM заключается в том, что он ищет не просто любую гиперплоскость, а ту, которая максимизирует расстояние (шумоподъемности) между ближайшими точками разных классов, называемыми опорными векторами.

Математическая основа

Формально задача SVM может быть представлена следующим образом: найти плоскость, которая разделяет классы, минимизируя ошибку классификации, и одновременно максимизируя расстояние между границами.

(2)

с условием, что:

(3)

где — вектор весов, — смещение, — признаки, а — метки классов.

Преимущества:

- эффективность в высокоразмерных данных: SVM хорошо работает в ситуациях, когда число признаков значительно превышает число наблюдений;

- гибкость: применение различных ядер (линейных, полиномиальных, радиальных) позволяет SVM решать задачи с нелинейными границами.

Недостатки:

- сложности с выбором параметров: Выбор ядра, а также параметры регуляризации могут значительно влиять на качество модели;

- время обучения: для очень больших выборок метод может быть вычислительно накладным.

## **1.3 Метод ближайших соседей (KNN)**

Метод ближайших соседей (K-Nearest Neighbors, KNN) — это простой, но мощный алгоритм, основанный на представлениях о сходстве. Основная идея KNN заключается в том, что класс нового объекта определяется по классам K ближайших объектов в обучающей выборке.

Математическая основа

Решение классификационной задачи с использованием KNN заключается в следующем:

1. Расчет расстояния между новым объектом и всеми объектами в обучающей выборке (обычно используется евклидово расстояние).

2. Выбор K ближайших соседей и голосование (для классификации) о принадлежности нового объекта к классу.

(4)

где I — индикаторная функция, — класс соседей, а C — множество всех классов.

Преимущества:

- простота и легкость в понимании: Алгоритм легко интерпретировать и использовать;

- отсутствие фазы обучения: Все данные используются для принятия решения, нет необходимости в предварительной подготовке модели.

Недостатки:

- время предсказания: для каждого нового объекта требуется вычислить расстояния до всех обучающих объектов, что может быть медленно для больших наборов данных;

- чувствительность к шуму: наличие выбросов или нерелевантных признаков может негативно влиять на точность классификации.

## **1.4 Случайный лес (Random Forest)**

Случайный лес является ансамблевым методом, который строится на основе нескольких деревьев решений. Каждое дерево обучается на случайной подвыборке обучающих данных, и его предсказания объединяются для получения финального решения. Этот алгоритм хорошо работает для задач классификации и регрессии, обеспечивая высокую точность и устойчивость.

Математическая основа

Суть метода заключается в построении множеств деревьев решений и их агрегации. Для классификации это можно выразить следующим образом:

где — предсказание i-го дерева, а n — общее количество деревьев в лесу.

Преимущества:

- высокая точность и стабильность: случайный лес значительно уменьшает риск переобучения по сравнению с одиночными деревьями решений;

- оценка значимости признаков: метод может легко оценить относительную важность различных признаков, что делает его полезным для анализа.

Недостатки:

- сложность интерпретации: хотя отдельные деревья легко интерпретируемы, ансамбль из множества деревьев становится сложным для анализа;

- затраты на ресурсы: Обучение большого количества деревьев может потребовать значительных вычислительных ресурсов, особенно для очень больших наборов данных.

Каждый из представленных методов классификации имеет свои особенности, преимущества и недостатки. Выбор подходящего алгоритма зависит от специфики задачи, доступных данных и требований к интерпретируемости. В дальнейшем исследовании будет применен один или несколько из этих методов для прогнозирования оттока клиентов фитнес-центров, что позволит выявить ключевые закономерности и повысить качество предсказаний.

# **2. ПРАКТИЧЕСКАЯ ЧАСТЬ. РЕАЛИЗАЦИЯ МОДЕЛЕЙ НА ДАННЫХ**

## **2.1. Знакомство и предобработка данных**

Сеть фитнес-центров «Культурист-датасаентист» разрабатывает стратегию взаимодействия с пользователями на основе аналитических данных.

Распространённая проблема фитнес-клубов и других сервисов — отток клиентов. Как понять, что пользователь больше не с вами? Можно записать в отток тех, кто попросил закрыть договор или удалил аккаунт. Однако клиенты не всегда уходят демонстративно: чаще перестают пользоваться сервисом тихо.

Индикаторы оттока зависят от специфики отрасли. Когда пользователь редко, но стабильно закупается в интернет-магазине — не похоже, что он «отвалился». А вот если две недели не заходит на канал с ежедневно обновляемым контентом, дела плохи: подписчик заскучал и, кажется, оставил вас.

Для фитнес-центра можно считать, что клиент попал в отток, если за последний месяц ни разу не посетил спортзал. Конечно, не исключено, что он уехал на Бали и по приезде обязательно продолжит ходить на фитнес. Однако чаще бывает наоборот. Если клиент начал новую жизнь с понедельника, немного походил в спортзал, а потом пропал — скорее всего, он не вернётся.

Чтобы бороться с оттоком, отдел по работе с клиентами «Культуриста-датасаентиста» перевёл в электронный вид множество анкет пользователей.

Наша задача: построить несколько моделей бинарной классификации пользователей, где целевой признак — факт оттока пользователя в следующем месяце, выбрать наилучшую модель основываясь на показателях метрик и провести оценку важностей признаков.

**Описание данных**

«Культурист-датасаентист» предоставил сведения в csv-файлах. Заказчик подготовил данные, которые содержат данные на месяц до оттока и факт оттока на определённый месяц. Набор данных включает следующие поля:

* Churn — факт оттока в текущем месяце;

Текущие поля в датасете:

Данные пользователя за предыдущий до проверки факта оттока месяц:

* gender — пол;
* Near\_Location — проживание или работа в районе, где находится фитнес-центр;
* Partner — сотрудник компании-партнёра клуба (сотрудничество с компаниями, чьи сотрудники могут получать скидки на абонемент — в таком случае фитнес-центр хранит информацию о работодателе клиента);
* Promo\_friends — факт первоначальной записи в рамках акции «приведи друга» (использовал промо-код от знакомого при оплате первого абонемента);
* Phone — наличие контактного телефона;
* Age — возраст;
* Lifetime — время с момента первого обращения в фитнес-центр (в месяцах).

Информация на основе журнала посещений, покупок и информация о текущем статусе абонемента клиента:

* Contract\_period — длительность текущего действующего абонемента (месяц, 3 месяца, 6 месяцев, год);
* Month\_to\_end\_contract — срок до окончания текущего действующего абонемента (в месяцах);
* Group\_visits — факт посещения групповых занятий;
* Avg\_class\_frequency\_total — средняя частота посещений в неделю за все время с начала действия абонемента;
* Avg\_class\_frequency\_current\_month — средняя частота посещений в неделю за предыдущий месяц;
* Avg\_additional\_charges\_total — суммарная выручка от других услуг фитнес-центра: кафе, спорт-товары, косметический и массажный салон.

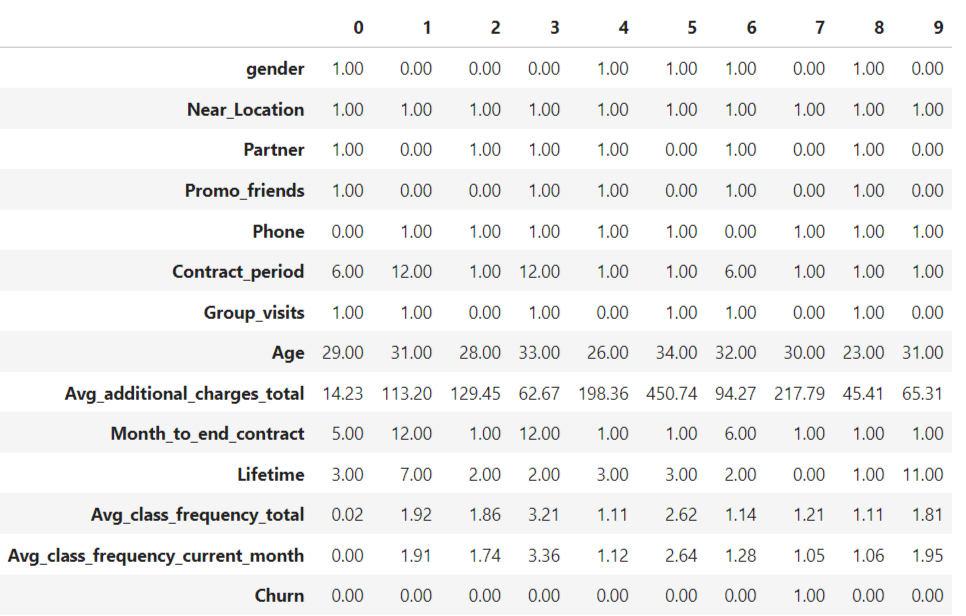


Рис. 1. Первые 10 строк датасета

**Предобработка данных**

Шаг 1. Проверка на дубликаты и пропуска показала, что ни того, ни того в данных нет (Приложение А);

Шаг 2. Убрали верхний регистр в названиях столбцов. (Приложение Б);

Шаг 3. Посмотрели на средние значения признаков в двух группах — тех, кто ушел в отток и тех, кто остался:

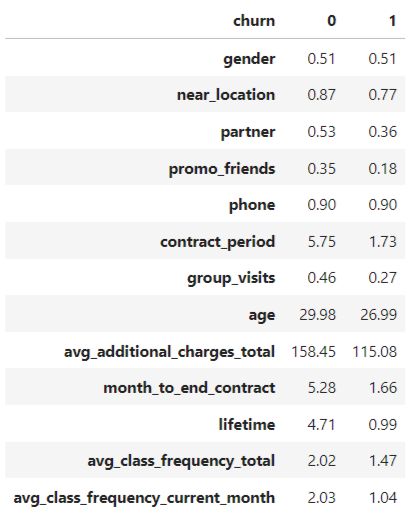


Рис. 2. Средние значения признаков в двух группах — тех, кто ушел в отток и тех, кто остался

Шаг 4. Выделили категориальные и количественные признаки.

В группу категориальных данных попали: gender, near\_location, partner, promo\_friends, phone, group\_visits, churn.

В группу количественных признаков попали: contract\_period, age, avg\_additional\_charges\_total, month\_to\_end\_contract, lifetime, avg\_class\_frequency\_total, avg\_class\_frequency\_current\_month.

Шаг 5. Построили столбчатые гистограммы и распределения признаков для тех, кто ушёл (отток) и тех, кто остался (не попали в отток);

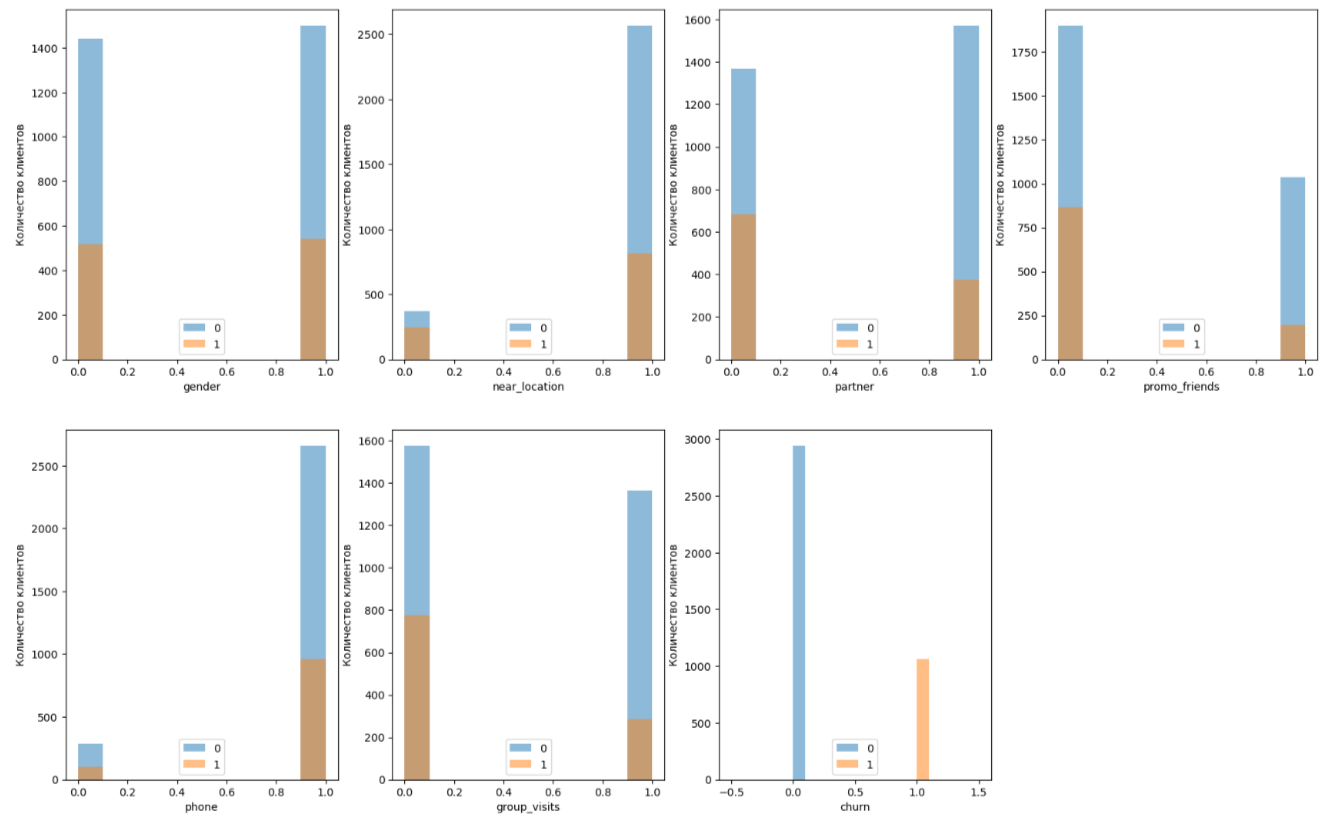
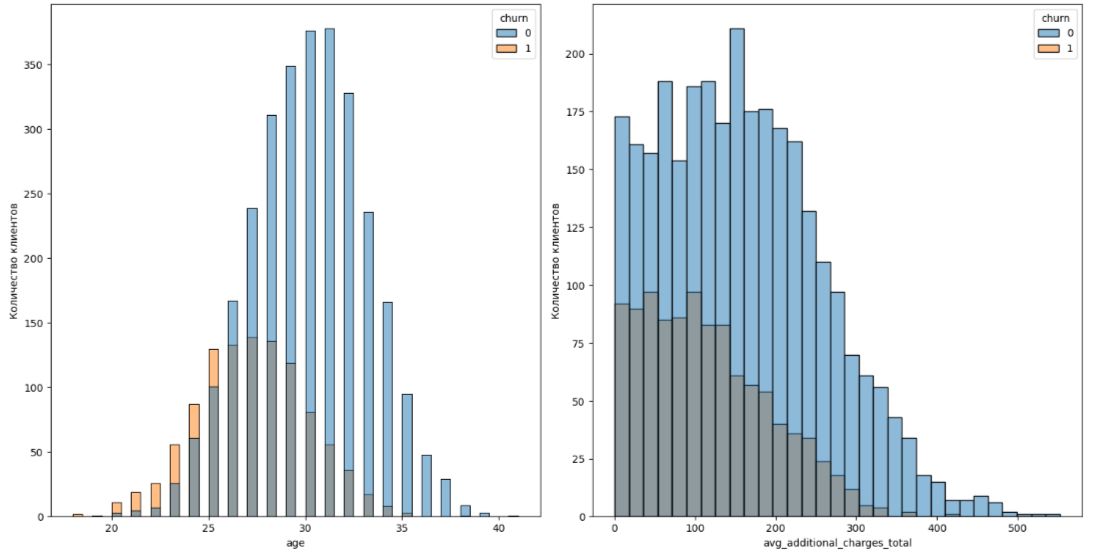


Рис. 3. Столбчатые гистограммы и распределения признаков для тех, кто ушёл (отток) и тех, кто остался (не попали в отток) для категориальных признаков.

Среди распределений бинарных признаков заметно, что чаще попадают в отток клиенты, которые проживают в другом районе, чей работодатель не сотрудничает с фитнес-центром и кто не ходит на групповые занятия.

Рис. 4.1. Столбчатые гистограммы и распределения признаков для тех, кто ушёл (отток) и тех, кто остался (не попали в отток) для количественных признаков (age и avg\_additional\_charges\_total)

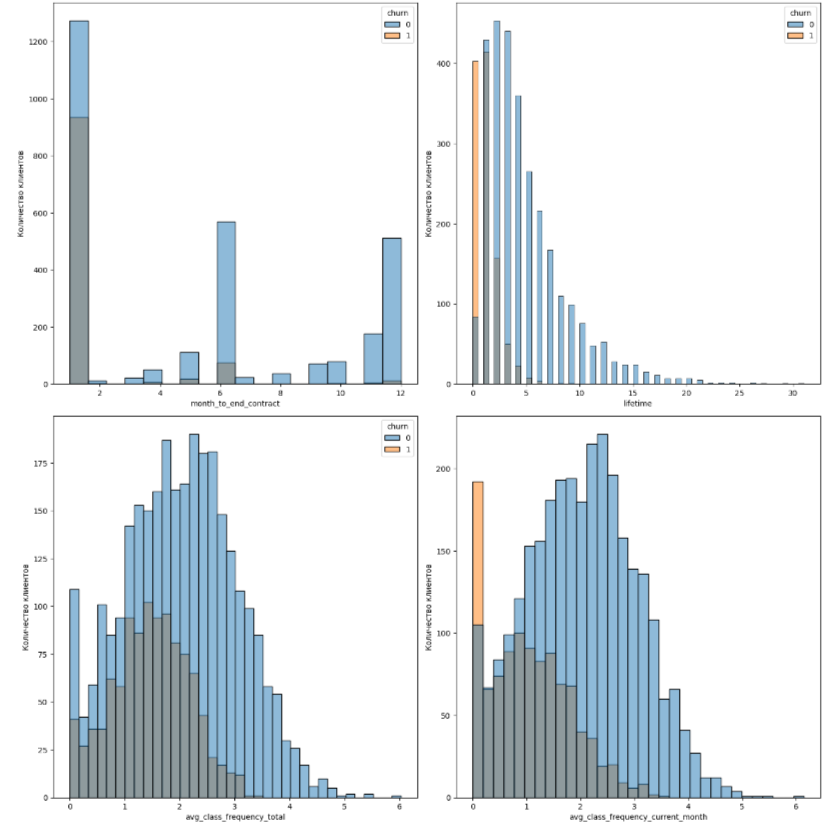


Рис. 4.2. Столбчатые гистограммы и распределения признаков - month\_to\_end\_contract, lifetime

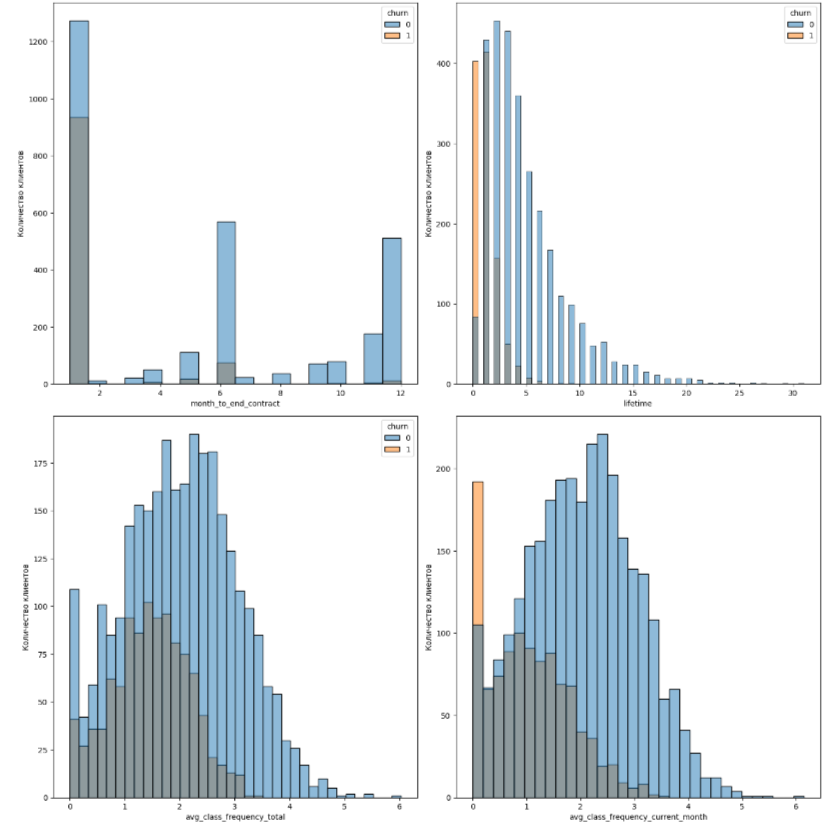


Рис. 4.2. Столбчатые гистограммы и распределения признаков -avg\_class\_frequency\_total, avg\_class\_frequency\_current\_month

Среди распределений небинарных признаков заметно, что чаще попадают в отток клиенты, которые не ходили в текущий месяц, купили абонемент всего на 1 месяц, у которых остался всего месяц абонемента, кто пришел впервые не более 3 месяцев назад. Также эти клиенты в среднем немного моложе и немного реже посещают тренировки в целом.

Шаг 6. Построили матриц корреляций для категориальных и количественных признаков, для проверки данных на мультиколлениарность.

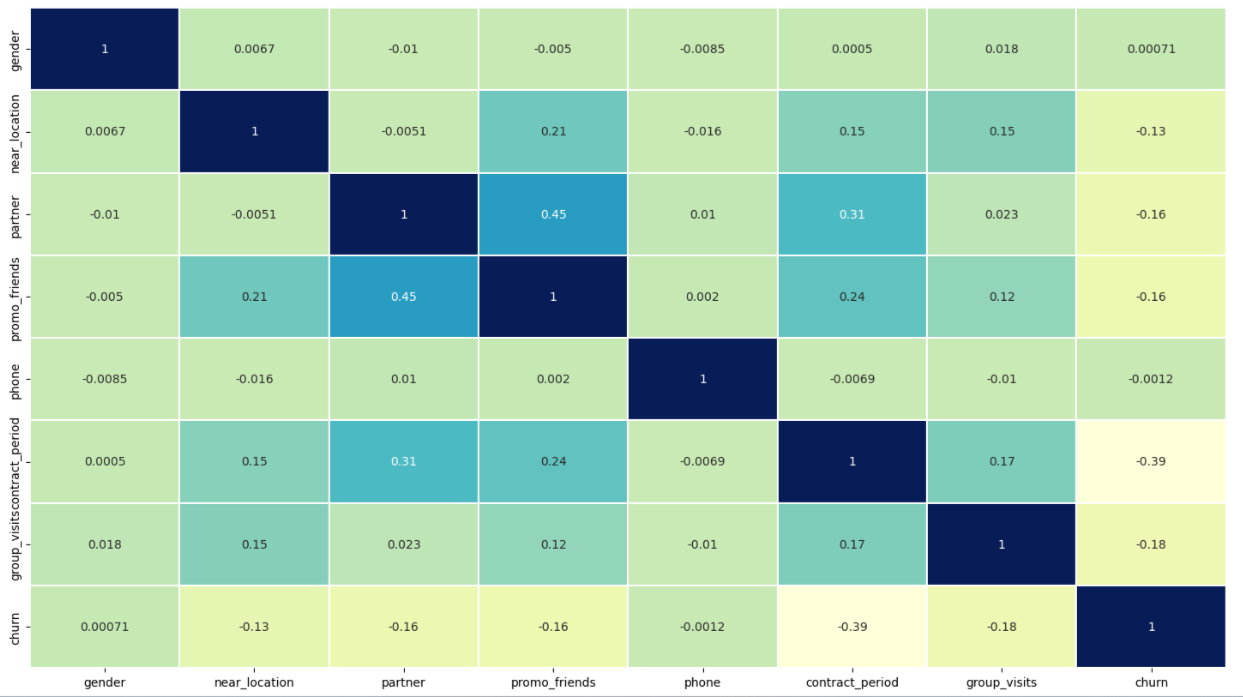


Рис. 5. Матрица корреляций для категориальных признаков

Низкие значения корреляции:

В этой матрице большинство значений корреляции находятся в диапазоне от -0.39 до 0.45, что указывает на отсутствие сильных взаимосвязей между признаками.

Наиболее заметная корреляция наблюдается между partner и promo\_friends (0.45), что может сигнализировать о наличии некоторой зависимости в поведении клиентов, но также не является критичной.

Негативные корреляции:

Корреляция между churn и age, а также churn и group\_visits показывает тенденцию к отрицательным значениям (например, -0.39 для churn и group\_visits). Это может означать, что более старшие клиенты или те, кто меньше участвует в групповых занятиях, склонны к большему оттоку.



Рис. 6. Матрица корреляций для количественных признаков

Сильные корреляции:

avg\_class\_frequency\_total и avg\_class\_frequency\_current\_month имеют очень высокую корреляцию (0.95). Это указывает на сильную мультиколлинеарность, что может оказать негативное влияние на модель, поскольку включение обоих признаков в одну модель может затруднить интерпретацию их отдельных эффектов.

Слабые положительные корреляции:

Прочие признаки (например, age, lifetime и month\_to\_end\_contract) демонстрируют слабые положительные корреляции (в диапазоне от 0.08 до 0.18). Эти связи, хотя и существуют, не являются сильными и могут не вызывать значительных проблем в модели.

Основной промежуточный вывод: Сильной прямой корреляции между оттоком нет ни с какими признаками.

## **2.2. Реализация моделей**

0. Предобработка данных

Определим препроцессор для данных, который включает преобразование числовых и категориальных признаков перед передачей данных в модель машинного обучения. Используем для этого колоннотрансформатор (ColumnTransformer), который объединяет преобразования числовых и категориальных признаков. (Приложение В)

1. Разделение данных на признаки (матрица Х) и целевую переменную (у). Формируем обучающую и тестовую выборки.

В качестве целевой переменной был выбран отток (1- клиент ушел, 0 -клиент остался). Генеральная совокупность была разделена на тестовую и обучающую выборки в пропорциях 20 на 80 соответственно. (Приложение Г)

X\_train: (3200, 13)

X\_test: (800, 13)

y\_train: (3200,)

y\_test: (800,)

2. Определяем функции для обучения и оценки моделей

На данном этапе прописываем функцию для обучения и оценки моделей, в которой происходит подбор гиперпараметров, время прогнозирования, сохранение лучших параметров и лучших кросс-валидируемых баллов и сохранение вероятностей для ROC-кривой.

Данная функция универсальна и далее используется для каждой модели. (Приложение Д)

3. Логистическая регрессия

Обучение и оценка логистической регрессии представлены в приложении Е. Также для логистической регрессии была построена ROC-кривая, представленная на рисунке 7.

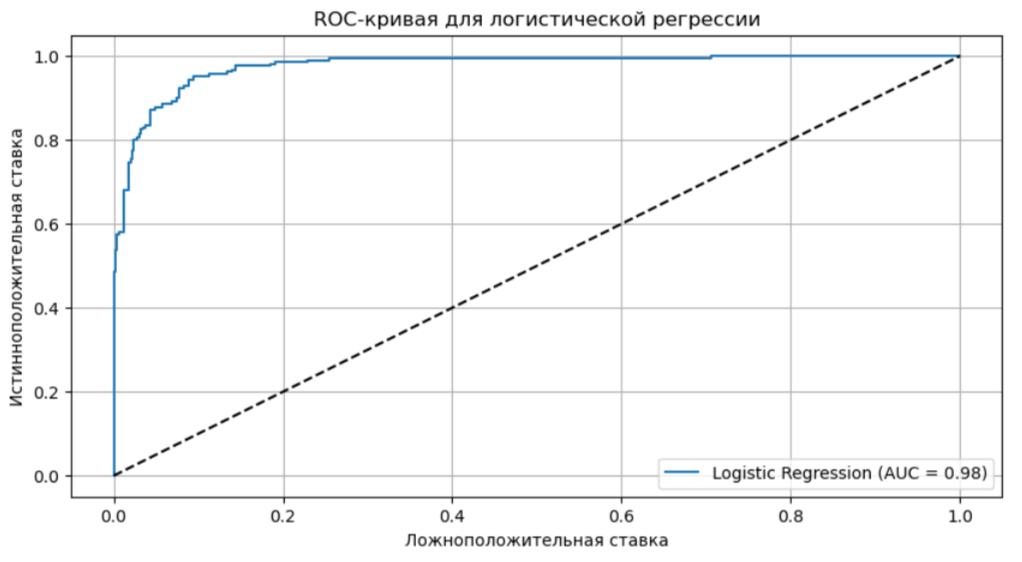
**

Рис. 7. ROC-кривая для логистической регрессии

Для логистической регрессии были получены следующие метрики:

- Accuracy: 0.93

- Precision: 0.92

- Recall: 0.90

- ROC AUC: 0.98

- Время прогнозирования: 0.00 секунд

- Лучшие параметры: {logregressor\_\_C: 0.1, logregressor\_\_penalty: 'l1'}

Обоснование выбора гиперпараметров:

- C: Этот параметр определяет степень регуляризации. Значение 0.1 подходит для обеспечения достаточной регуляризации, что уменьшает переобучение, особенно в задачах с высокой размерностью.

- penalty: Выбор L1-пенальти добавляет отсечённость, что может привести к выбору наиболее значимых признаков за счёт зануления менее значимых. Это особенно полезно при наличии большого количества признаков, когда некоторые могут быть лишними.

4. Метод опорных векторов (SVM)

Обучение и оценка метода опорных векторов представлены в приложении Ж. Также для логистической регрессии была построена ROC-кривая, представленная на рисунке 8.

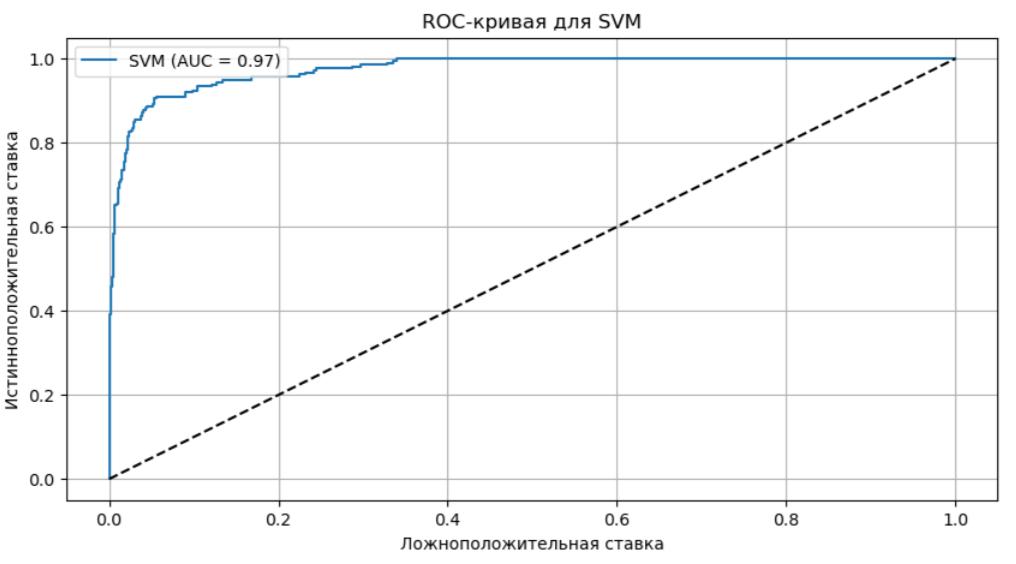
**

Рис. 8. ROC-кривая для метода опорных векторов.

Для метода опорных векторов были получены следующие метрики:

- Accuracy: 0.93

- Precision: 0.92

- Recall: 0.91

- ROC AUC: 0.97

- Время прогнозирования: 0.09 секунд

- Лучшие параметры:{svm\_\_C: 10, svm\_\_gamma: 'auto', svm\_\_kernel: 'rbf'}

Обоснование выбора гиперпараметров:

- C: Значение `10` выбрано для обеспечения меньшей регуляризации и увеличения сложности модели, что позволяет захватывать более сложные паттерны в данных. Это может быть полезно, если данные не сильно шумные.

- gamma: Параметр `auto` автоматически устанавливает значение gamma в зависимости от количества признаков. Верифицированные паттерны более эффективно обрабатываются с учетом нелинейной зависимости.

- kernel: Ядро RBF (радиально-базисная функция) используется для обеспечения гибкости в моделировании данных с ненормальными границами классов, что делает SVM мощным инструментом для неклассических задач классификации.

5. Метод ближайших соседей (KNN)

Обучение и оценка метода ближайших соседей представлены в приложении З. Также для логистической регрессии была построена ROC-кривая, представленная на рисунке 9.

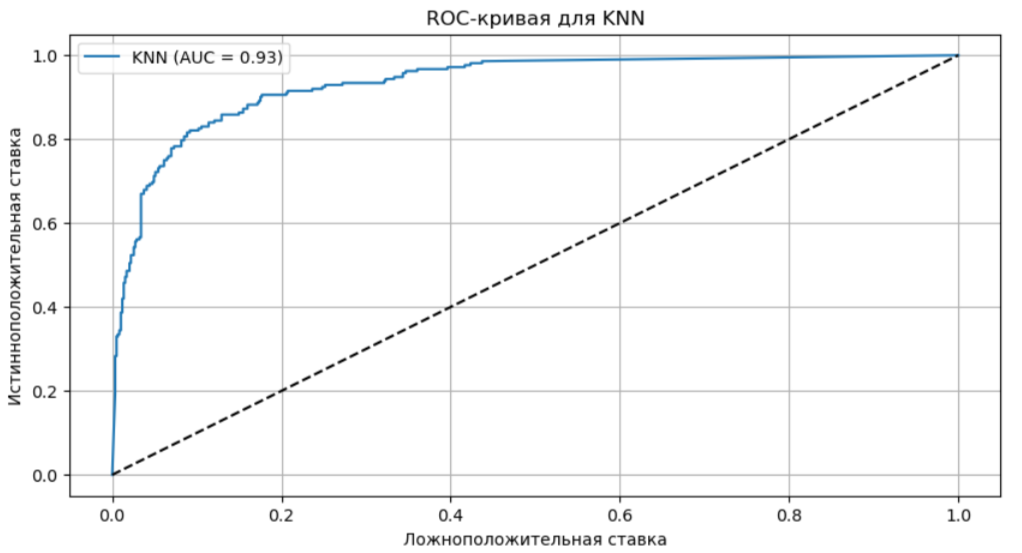


Рис. 9. ROC-кривая для метода ближайших соседей.

Для метода ближайших соседей были получены следующие метрики:

- Accuracy: 0.89

- Precision: 0.86

- Recall: 0.85

- ROC AUC: 0.93

- Время прогнозирования: 0.09 секунд

- Лучшие параметры: {knn\_\_metric: 'euclidean', knn\_\_n\_neighbors: 11, knn\_\_weights: 'distance'}`

Обоснование выбора гиперпараметров:

- n\_neighbors: Значение `11` оптимально, так как это число соседей позволяет достигнуть устойчивой классификации, минимизируя шум, что было доказано в процессе обучения.

- weights: Взвешивание по расстоянию позволяет моделировать так, что ближние объекты имеют больший вес в определении класса, что калькулирует более точные обстоятельства.

- metric: Использование евклидова расстояния является стандартным выбором в задачах с непрерывными признаками, что помогает обеспечить сопоставимость между точками.

6. Случайный лес (Random Forest)

Обучение и оценка метода ближайших соседей представлены в приложении И. Также для логистической регрессии была построена ROC-кривая, представленная на рисунке 10.

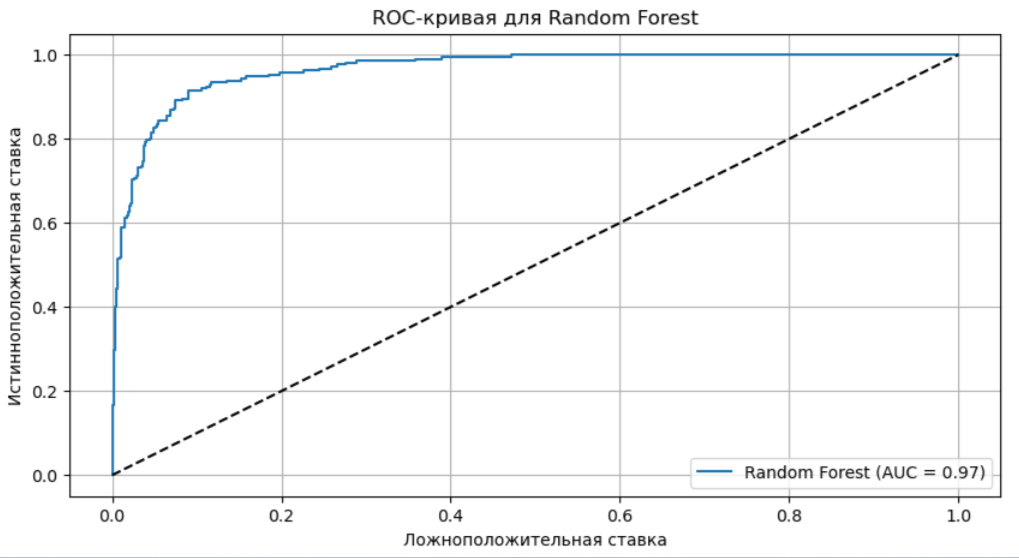


Рис. 10. ROC-кривая для случайного леса.

Для случайного леса были получены следующие метрики:

- Accuracy: 0.91

- Precision: 0.90

- Recall: 0.87

- ROC AUC: 0.97

- Время прогнозирования: 0.01 секунд

-Лучшие параметры: {randomforest\_\_max\_depth: 10, randomforest\_\_min\_samples\_split: 10, randomforest\_\_n\_estimators: 50}

Обоснование выбора гиперпараметров:

- max\_depth: Значение `10` определяет глубину дерева, что позволяет избежать переобучения, поддерживая при этом достаточную сложность.

- min\_samples\_split: Параметр `10` указывает минимальное количество образцов для разделения узла, что помогает уменьшить вариативность при обучении.

- n\_estimators: Выбор `50` деревьев в лесу обеспечивает хорошее качество предсказания при меньших затратах по времени.

7. Вывод результатов, матриц ошибок и общего графика ROC AUC

Общие результаты метрик по каждой модели представлены в таблице 1. С кодом для вывода общего графика ROC AUC и матриц ошибок по каждой модели можно ознакомиться в приложении К.

Таблица 1. Результаты метрик по каждой модели.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель\Метрика | **Accuracy** | **Precision** | **Recall** | **ROC AUC** |
| Logistic Regression | 0.93 | 0.92 | 0.90 | 0.98 |
| SVM | 0.93 | 0.92 | 0.91 | 0.97 |
| KNN | 0.89 | 0.86 | 0.85 | 0.93 |
| Random Forest | 0.91 | 0.90 | 0.87 | 0.97 |

Лучший кросс-валидируемый балл

1. Logistic Regression: 0.9290625

2. SVM: 0.9334375000000001

3. KNN: 0.8956250000000001

4. Random Forest: 0.9162500000000001

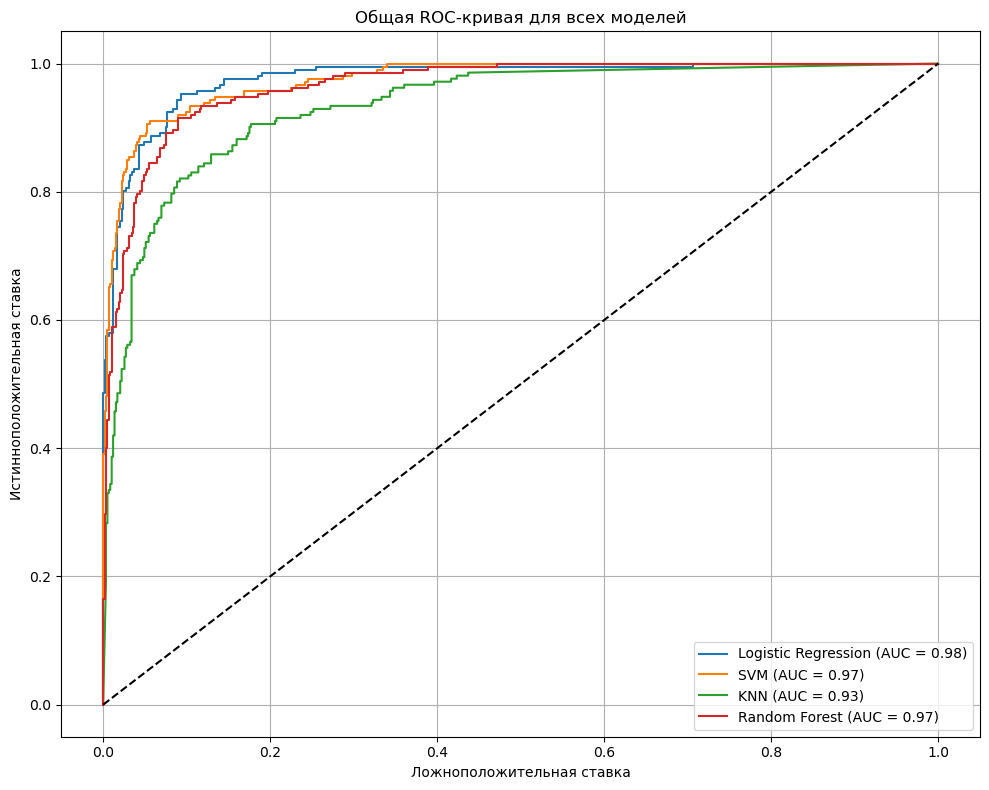


Рис.11. Общая ROC-кривая для всех моделей

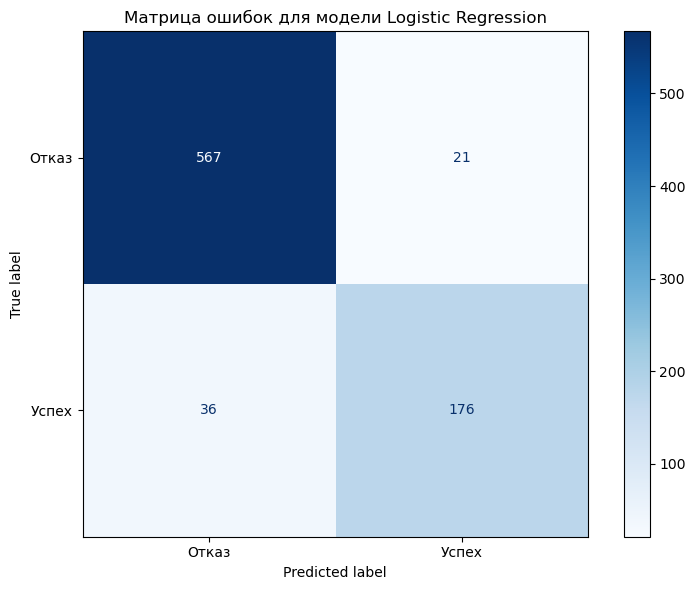


Рис.12. Матрица ошибок для логистической регрессии

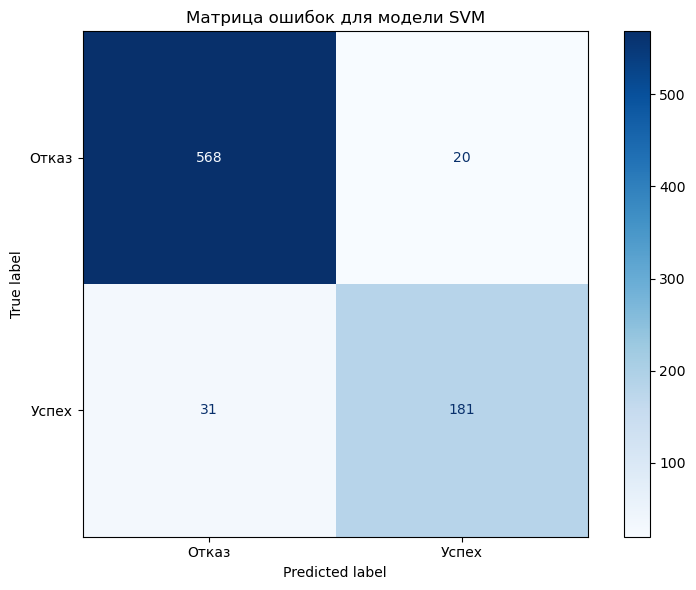


Рис.13. Матрица ошибок для метода опорных векторов (SVM).

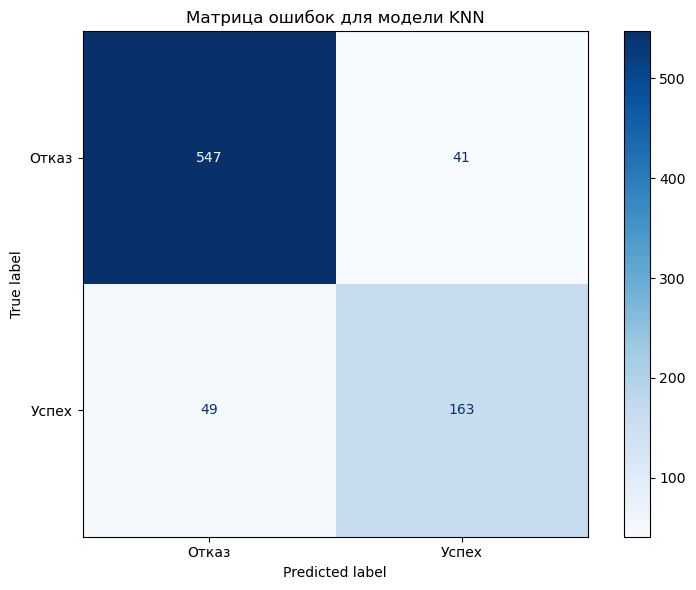


Рис.14. Матрица ошибок для метода ближайших соседей (KNN).

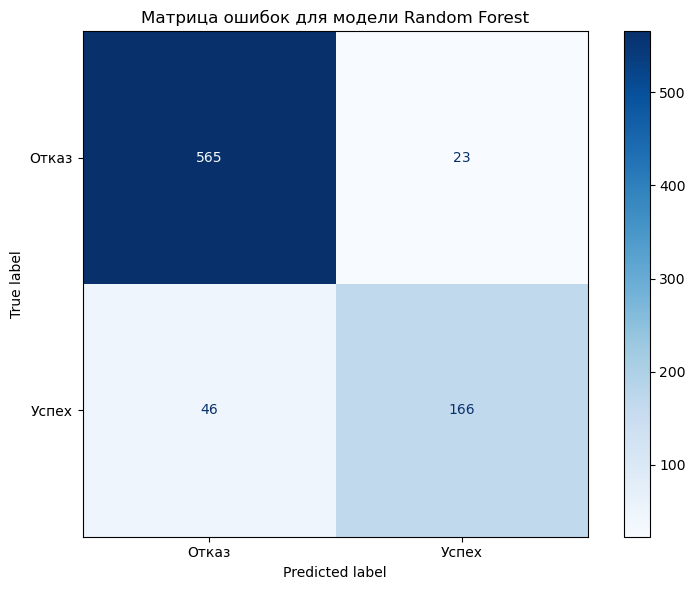


Рис.15. Матрица ошибок для случайного леса (Random Forest).

*Логистическая регрессия* демонстрирует высокую точность и устойчивость, особенно в метрике ROC AUC, что указывает на её способность эффективно различать классы. Это также подтверждается высоким кросс-валидировочным баллом, что делает её надежным выбором для данной задачи.

*Метод опорных векторов (SVM)* показывает схожую по сравнению с логистической регрессией точность, и даже немного превосходит её в кросс-валидации. Высокая полнота (0.91) также предполагает, что модель хорошо справляется с идентификацией положительных классов. Метод опорных векторов является сильным кандидатом в данной задаче.

*Метод ближайших соседей (KNN*) имеет наименьшие показатели среди всех моделей, что указывает на ограничения этого метода в изучаемой задаче. Низкое значение точности и полноты говорит о том, что KNN может не быть готовым к заявленному уровню сложности данных.

*Случайный лес (Random Forest)* демонстрирует хорошие результаты, аналогичные SVM, но все же уступает в кросс-валидации. Несмотря на высокую точность и полноту, его результаты ниже, чем у SVM в рамках кросс-валидации.

Сравнительный анализ моделей

- *Качество предсказания*: наилучшие показатели точности и AUC показаны у логистической регрессии. SVM также демонстрирует высокие результаты, но время прогнозирования у обеих моделей минимально;

- *Скорость прогнозирования*: все модели выполняются быстро, однако логистическая регрессия оказывается наиболее быстрой, что делает её предпочтительным вариантом в условиях ограниченных ресурсов;

- *Интерпретируемость*: логистическая регрессия имеет наивысшую интерпретируемость, поскольку коэффициенты могут напрямую быть связаны с вероятностями классов. Mетод опорных векторов и случайный лес менее интерпретируемы; однако, правая динамика для KNN среди всех кажется самой низкой из-за "чёрного ящика";

- *Визуализация*: логистическая регрессия и KNN наиболее просто визуализируются из-за их простой структуры и глубины. Случайный лес является более сложным в этом аспекте.

В качестве лучшей модели для следующего пункта выберем логистическую регрессию. Для нее и выведем топ-5 самых значимых признаков.

## **2.3 Оценка важности признаков**

Для логистической регрессии был выведен график важности признаков (топ-5) на рисунке 16.

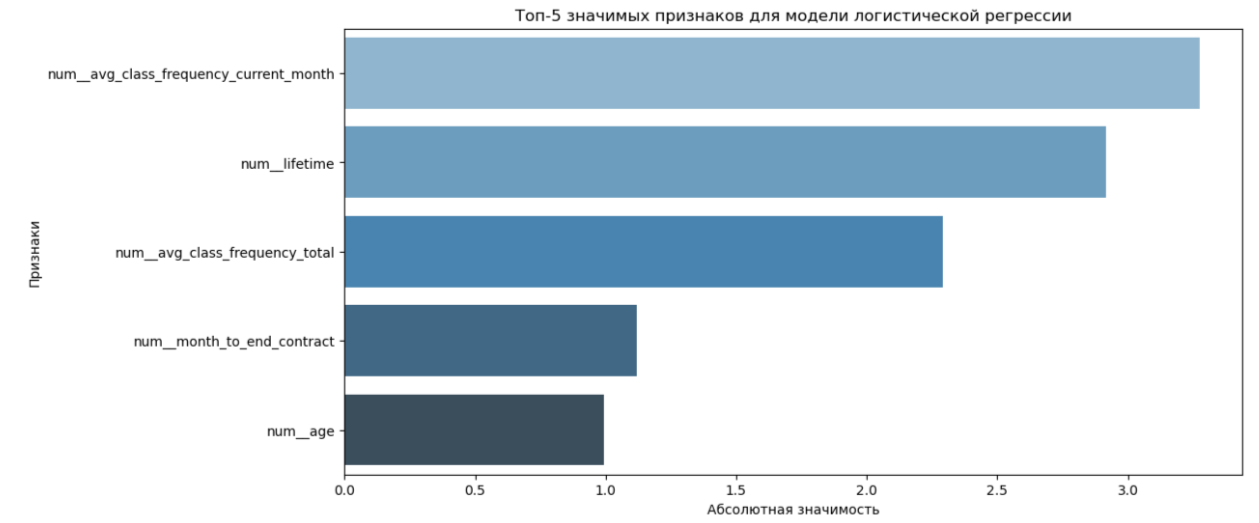


Рис. 16. Топ-5 значимых признаков для модели логической регрессии.

На графике представлены топ-5 самых значимых признаков для модели логистической регрессии в контексте прогнозирования оттока клиентов (churn) в фитнес-центре. Рассмотрим каждый из выделенных признаков:

1. Средняя частота посещений в текущем месяце (avg\_class\_frequency\_current\_month):

Этот признак имеет наивысшую абсолютную значимость, что указывает на то, что частота посещений клиентов в последние недели является ключевым фактором, влияющим на вероятность их оттока. Более высокая частота посещений, как правило, коррелирует с удовлетворенностью клиентов и их желанием продолжать абонемент.

1. Время с момента первого обращения (lifetime):

Чем дольше клиент пользуется услугами фитнес-центра, тем ниже вероятность его ухода, что указывает на высокую лояльность к бренду.

1. Средняя частота посещений за все время (avg\_class\_frequency\_total):

Клиенты, посещающие фитнес-центр часто, с большей вероятностью останутся, так как получают максимальную выгоду от своих абонементов.

1. Срок до окончания текущего действующего абонемента (month\_to\_end\_contract):

Близость к окончанию срока действия абонемента повышает вероятность оттока, так как клиенты начинают принимать решение о продлении.

1. Возраст (age):

Хотя и менее значимо, чем другие факторы, возраст клиента также влияет на вероятность оттока. Молодые клиенты более склонны к перемене фитнес-центров, в отличие от более зрелых.

## **2.4. Вывод результатов и их анализ**

Выбор лучшей модели зависит от конкретных требований задачи. Если задача требует высокой интерпретируемости и скорости, **логистическая регрессия** будет наилучшим выбором. **Метод опорных векторов** предлагает гибкость и обнаружение сложных зависимостей, что делает его сильным кандидатом при наличии больших объемов данных. **Случайный лес** рекомендуется для задач, где стабильность и качество важнее интерпретируемости. **KNN** может обеспечить стабильные результаты на небольших наборах данных, но его масштабируемость ограничена.

Анализ важности признаков для модели логистической регрессии показывает, что частота посещений в текущем месяце и продолжительность времени с момента первого обращения к фитнес-центру являются наиболее значимыми показателями, влияющими на отток клиентов.

# **3. ОБЩИЕ ВЫВОДЫ И РЕКОМЕНДАЦИИ**

Анализ данных о клиентах фитнес-центра, включая результаты модели логистической регрессии и важность различных признаков, позволяет сделать несколько ключевых выводов и предложить практические рекомендации, направленные на снижение оттока клиентов и улучшение их удержания.

## **3.1 Важность частоты посещений**

Средняя частота посещений клиентами фитнес-центра в течение месяца сыграла ключевую роль в их решении остаться в фитнес-центре или покинуть его. Высокая частота посещений тесно связана с общим уровнем удовлетворенности предоставляемыми услугами и вовлеченностью клиентов.

Рекомендации:

-Разработка программ лояльности: внедрить бонусную систему для предоставления клиентам скидок или бесплатных курсов при достижении ими определенного уровня посещаемости (например, 10 раз в месяц);

-Организация мероприятий и промо-акций: регулярно проводите специальные занятия, конкурсы и групповые тренинги, чтобы побудить клиентов чаще посещать центр;

-Персонализированные предложения: Используйте функцию анализа для создания индивидуальных предложений на основе вашей истории посещений, чтобы повысить вовлеченность клиентов.

## **3.2 Продолжительность времени с момента первого обращения**

Клиенты, имеющие более длительный опыт взаимодействия с фитнес-центрами, вряд ли будут потеряны, что подчеркивает важность построения долгосрочных отношений.

Рекомендации:

-План приветствия: запустите специальный план для новых участников, который поможет им адаптироваться и получить положительный опыт в первые несколько месяцев;

-Регулярное общение: поддерживайте связь с клиентами с помощью электронной почты, SMS или мобильных приложений, чтобы информировать их о новых предложениях, рекламных акциях и предстоящих событиях.

## **3.3 Срок до окончания действующего абонемента**

Признак, показывающий срок до окончания текущего абонемента, также является критически важным для прогнозирования оттока. Чем ближе срок окончания, тем выше вероятность, что клиент размышляет о прекращении абонемента.

*Рекомендации*:

- Заранее информируйте клиентов: Напоминайте клиентам о предстоящем окончании абонемента за несколько недель до даты окончания, предлагая специальные условия для продления.

- Стимулируйте продление: Предлагайте скидки или бонусы за продление абонемента за несколько месяцев вперед, чтобы повысить лояльность и снизить уровень оттока.

## **3.4 Возраст клиентов**

Возраст клиентов является значимым фактором, влияющим на их решения о продолжении посещения фитнес-центра.

*Рекомендации*:

- Диверсификация программ: Разработайте разнообразные программы тренировок, учитывая разные возрастные группы. Например, группы для молодежи, программы для людей среднего возраста и специализированные занятия для пожилых клиентов.

- Открытие возрастных клубов: Создайте специализированные клубы или группы, которые ориентированы на определенные возрастные категории, что поможет клиентам чувствовать себя более комфортно и вовлеченно.

## **3.5 Использование дополнительного дохода**

Суммарная выручка от дополнительных услуг (кафе, спорт-товары) также может оказывать влияние на удержание клиентов. Удовлетворенность дополнительными услугами может повысить общий уровень удовлетворенности.

*Рекомендации*:

- Улучшение дополнительных услуг: Анализируйте предпочтения клиентов и улучшайте предлагаемые услуги — например, расширение меню в кафе, введение новых услуг в спа-центре и т.д.

- Кросс-промоции: Предлагайте специальные акции, когда клиенты пользуются услугами фитнеса вместе с дополнительными услугами, что может увеличить общую выручку и удержание клиентов.

# **ЗАКЛЮЧЕНИЕ**

Анализ данных о клиентах фитнес-центра выявил важные факторы, влияющие на отток клиентов и их лояльность. Основное внимание следует уделить частоте посещений, которая служит критерием вовлеченности. Высокая частота коррелирует с удовлетворенностью услугами, что открывает возможности для разработки мотивационных программ, направленных на увеличении посещаемости.

Продолжительность времени пребывания клиента в фитнес-центре также играет ключевую роль. Долгосрочные отношения позволяют формировать доверие и лояльность. Приветственные программы для новых клиентов должны стать стандартной практикой, чтобы обеспечить положительный опыт и уменьшить риск оттока.

Важно также активное управление отношениями перед окончанием срока абонемента. Прозрачные коммуникации о продлении и привлекательные предложения могут значительно снизить вероятность ухода клиентов. Поддержание постоянного контакта поможет повысить уровень удовлетворенности и вовлеченности.

Наконец, предложение дополнительных услуг, таких как кафе или секции с оборудованием, может влиять на retention. Повышение качества этих услуг сформирует положительное впечатление и сделает фитнес-центр не только местом для тренировок, но и пространством для отдыха и общения, что в целом улучшает общий клиентский опыт.

# **Список использованной литературы**

1. Документация scipy логистическая регрессия [Электронный источник] / URL: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html> (дата обращения 03.12.2024).

2. Документация scipy метод опорных векторов [Электронный источник] / URL: <https://scikit-learn.ru/stable/modules/svm.html> (дата обращения 04.12.2024).

3. Документация scipy метод ближайших соседей [Электронный источник] / URL: <https://scikit-learn.ru/stable/modules/neighbors.html> (дата обращения 03.12.2024).

4. Документация scipy случайный лес [Электронный источник] / URL: <https://scikit-learn.ru/stable/modules/ensemble.html> (дата обращения 05.12.2024).

5. Документация библиотеки matplotlib [Электронный источник] / URL: <https://matplotlib.org/stable/index.html> (дата обращения 07.12.2024).

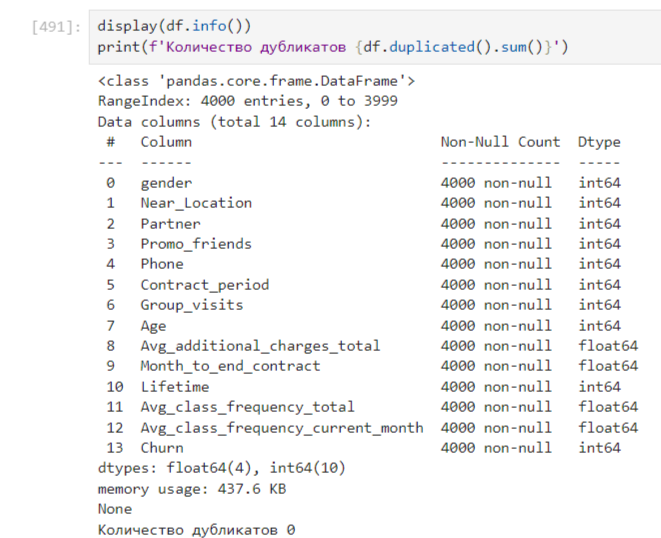
6. Курс от Яндекс Практикум «Аналитик данных». Спринт «Основы машинного обучения» - источник данных.

7. Яндекс образование / хэнд бук «Машинное обучение» [Электронный источник] / URL: <https://education.yandex.ru/handbook/ml> (дата обращения 08.12.2024).

8. SVM. Подробный разбор метода опорных векторов, реализация на python [Электронный источник] / URL: <https://habr.com/ru/companies/skillfactory/articles/701530/> (дата обращения 08.12.2024).

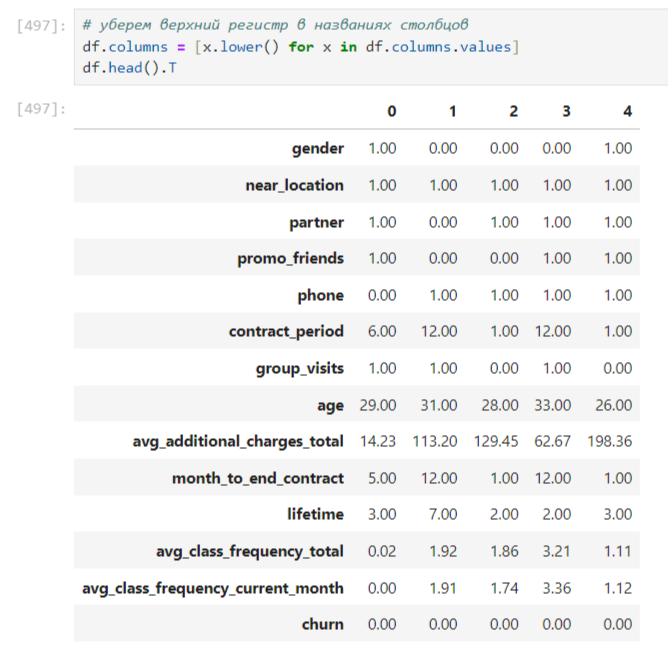
## ПРИЛОЖЕНИЕ А

## Проверка на пропуски и дубликаты



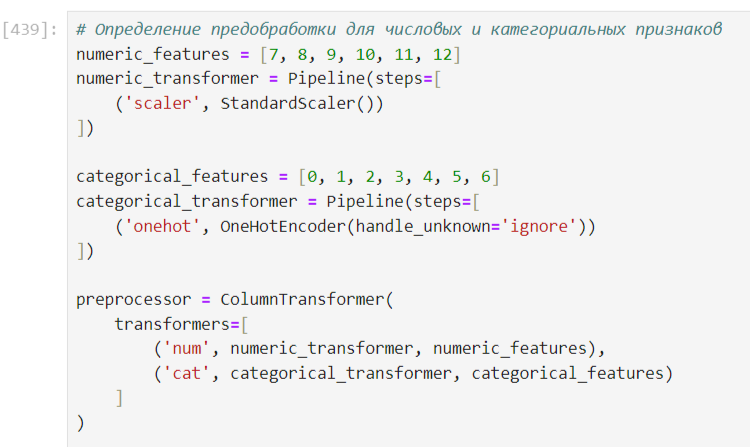
## ПРИЛОЖЕНИЕ Б

## Убрали верхний регистр столбцов



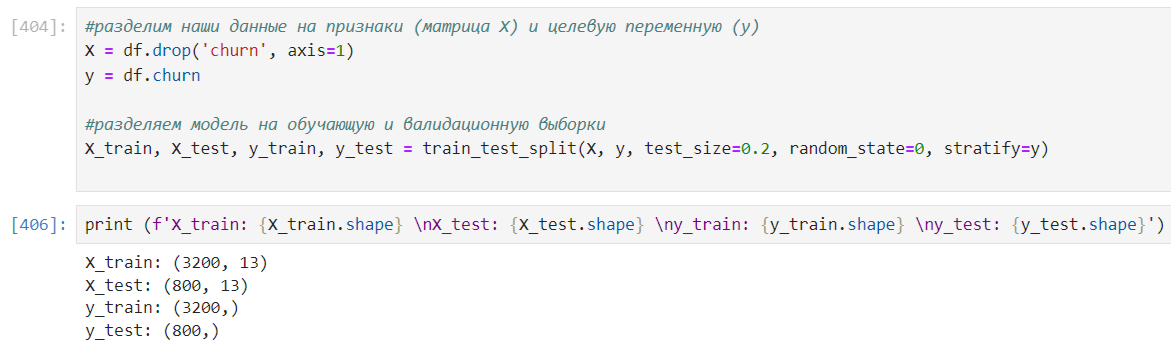
## ПРИЛОЖЕНИЕ В

## Предобработка данных



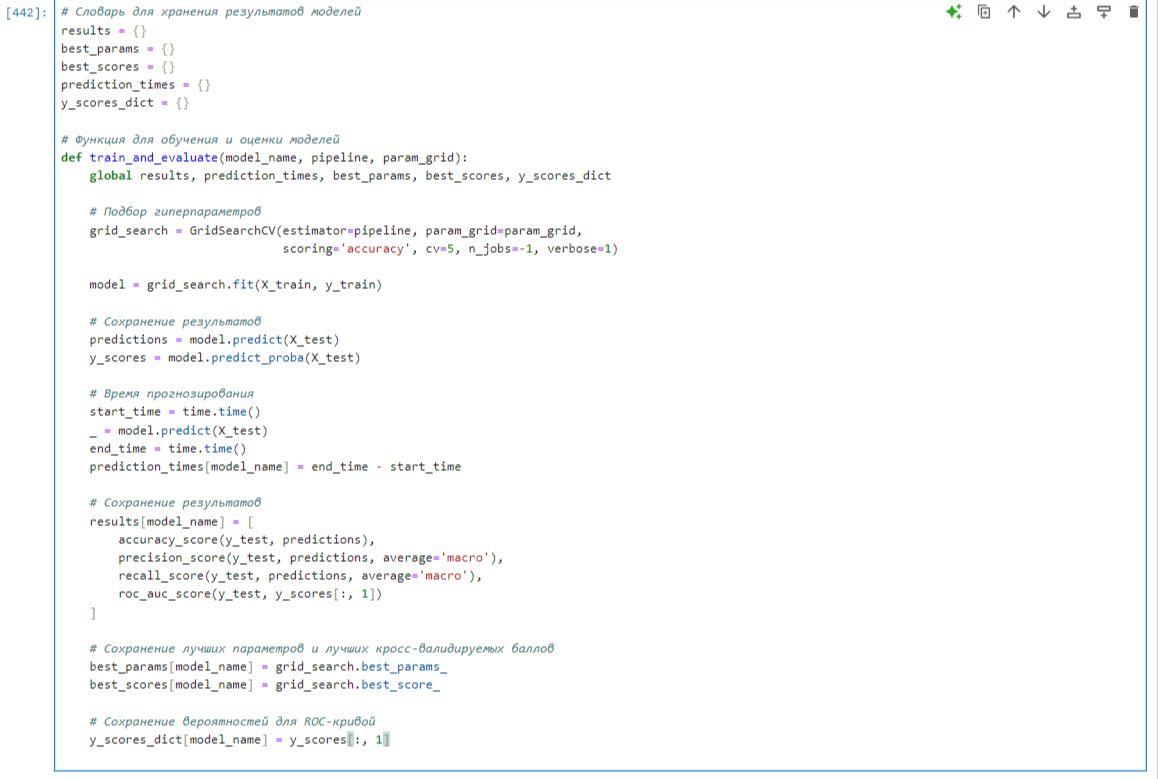
## ПРИЛОЖЕНИЕ Г

## Разделение данных на признаки (матрица Х) и целевую переменную (у). Формируем обучающую и тестовую выборки.



## ПРИЛОЖЕНИЕ Д

## Определяем функции для обучения и оценки моделей



## ПРИЛОЖЕНИЕ Е

## Логистическая регрессия



## ПРИЛОЖЕНИЕ Ж

## Метод опорных векторов



## ПРИЛОЖЕНИЕ З

## Метод ближайших соседей



## ПРИЛОЖЕНИЕ И

## Случайный лес



## ПРИЛОЖЕНИЕ К

## Общий график ROC AUC и матрицы ошибок для каждой модели

