МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий (ИИТ)
Кафедра прикладной математики (ПМ)

КУРСОВАЯ РАБОТА

по дисциплине «Технологии организации обработки и хранения статистических данных»

Тема курсовой работы: «Прогнозирование риска сердечного приступа»

Студент группы ИМБО-02-22	Лищенко Тимофей Викторович	_____ (подпись)
Руководитель курсовой работы	доцент, к.п.н. Митина О.А.	_____ (подпись)
Работа представлена к защите	«___»_____2023 г.	
Допущен к защите	«___»_____2023 г.	

Москва 2023 г.

МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий (ИИТ)
Кафедра прикладной математики (ПМ)

Утверждаю
Заведующий кафедрой ПМ
_____Смоленцева Т.Е.
(подпись)
«21» сентября 2023 г.

ЗАДАНИЕ
на выполнение курсовой работы

по дисциплине «Технологии организации обработки и хранения статистических данных»

Студент Лищенко Тимофей Викторович Группа ИМБО-02-22

Тема «Прогнозирование риска сердечного приступа»

Исходные данные: выбранный студентом данные.
Перечень вопросов, подлежащих разработке, и обязательного графического материала:
Характеристика и общее описание признаков, влияющих на развитие сердечного приступа (описание признаков, типов данных, общее текстовое описание сердечных приступов и графическое представление влияния признаков на риск)
Графические модели зависимости влияния признаков на риск развития сердечного приступа (описание путем представления графиков по признакам)
Анализ риска возникновения сердечного приступа (выявление зависимостей между признаками)
Прогнозирование риска сердечного приступа.
Срок представления к защите курсовой работы: **до** «15» декабря 2023 г.

Задание на курсовую работу выдал		Митина О.А.
	Подпись руководителя	*(ФИО руководителя)*
		«21» сентября 2023 г.
Задание на курсовую работу получил		Лищенко Т.В.
	Подпись обучающегося	*(ФИО обучающегося)*
		«21» сентября 2023 г.

Москва 2023 г.

ОТЗЫВ

на курсовую работу

по дисциплине «Технологии организации, обработки и хранения статистических данных»

Студент Лищенко Тимофей Викторович **ИМБО-02-20**

(ФИО студента) (Группа)

Характеристика курсовой работы

Критерий	Да	Нет	Не полностью
1. Соответствие содержания курсовой работы указанной теме			
2. Соответствие курсовой работы заданию			
3. Соответствие рекомендациям по оформлению текста, таблиц, рисунков и пр.			
4. Полнота выполнения всех пунктов задания			
5. Логичность и системность содержания курсовой работы			
6. Отсутствие фактических грубых ошибок			

Замечаний:

Рекомендуемая оценка:

_____ _____

Подпись руководителя ФИО руководителя

Москва 2023 г.

Оглавление

ВВЕДЕНИЕ

В современном мире люди все чаще подвержены риску развития сердечного приступа, что требует всестороннего изучения связей между биологическими признаками человека и риском возникновения инфаркта.

При здоровом образе жизни все органы и системы органов человека, в частности кровеносная и сердечно-сосудистая системы, функционируют исправно, что минимизирует риск развития сердечного приступа. Поскольку все системы взаимосвязаны изменение в одной системе провоцирует неправильное функционирование всего организма. Важно оценивать то, насколько состояние здоровья, возраст, вредные привычки, ведение неправильного образа жизни влияют или связаны с риском возникновения сердечного приступа.

Таким образом, тема данной курсовой работы является актуальной для медицины и человека в целом.

Целью данной курсовой работы является – провести анализ и построить модель прогнозирования риска сердечного приступа.

Задачи, решаемые в данной курсовой работе:

1. Изучение научной и методической литературы о сердечно-сосудистых заболеваниях, а именно сердечных приступах, таких как инфаркт;
2. Поиск/Сбор данных о сердечных приступах;
3. При необходимости подготовить данные, провести очистку и предобработку;
4. Проанализировать исходные данные, построив графики зависимостей (провести первичный анализ данных);
5. Используя знания математической статистики и современных средств обработки данных: язык программирования Python с его библиотеками, построить модель прогнозирования.

1 Теоретическая часть

1.1 Сердечный приступ

Перед тем как провести анализ риска сердечного приступа, давайте попробуем разобраться в том, что представляет собой сердечный приступ. Сердечный приступ — это серьезное патологическое состояние, вызванное резким нарушением кровоснабжения сердечной мышцы, чаще всего вызванное закупоркой артерии тромбом и/или спазмом, обычно в месте атеросклеротического нарушения артерии, которая снабжает сердце кровью, что влечет за собой развитие ишемии и некроза (отмирания) участка сердечной мышцы.

Основные признаки сердечного приступа включают в себя следующие:

1. Боль в груди или дискомфорт.
2. Боль в груди или дискомфорт могут ощущаться как давление, стеснение или ощущение сдавливания.
3. Одышка: затрудненное дыхание, которое может сопровождаться болью в груди или происходить без нее.
4. Дискомфорт в других частях тела, таких как спина, руки, плечи, шея или челюсть.
5. Нарушение ритма сердца.

Смертность от острого инфаркта миокарда составляет около 30%, и еще 5-10% выживших после инфаркта миокарда умирают в течение года после него. Кроме того, после инфаркта миокарда человек может испытывать стресс, тревогу и депрессию. Даже легкие формы депрессии наблюдаются у 2/3 пациентов, находившихся в больнице из-за острого инфаркта миокарда, а тяжелая депрессия у примерно 15% всех пациентов с сердечно-сосудистыми заболеваниями. Это в несколько раз выше, чем в среднем в населении. Известно также, что депрессия увеличивает риск неблагоприятного исхода: у пациентов с депрессией, развившейся после инфаркта миокарда, риск смерти увеличивается в три раза.

Факторы риска инфаркта миокарда делятся на модифицируемые (на которые можно влиять) и не модифицируемые (которые нельзя изменить). Совокупность нескольких факторов риска увеличивает вероятность инфаркта миокарда и других сердечно-сосудистых заболеваний.

Модифицируемые факторы:

Курение повреждает кровеносные сосуды, увеличивая риск атеросклероза и образования тромбов. Даже пассивное курение увеличивает риск сердечно-сосудистых заболеваний.

Высокое артериальное давление (гипертония) повышает нагрузку на сердце, повреждает артерии, снабжающие сердце кровью, и увеличивает риск развития сердечной и почечной недостаточности, а также инфаркта миокарда и инсульта.

Высокий уровень холестерина в организме способствует образованию атеросклеротических бляшек на стенках сосудов.

Избыточный вес является фактором риска для высокого артериального давления, уровня холестерина, метаболического синдрома, сахарного диабета и других заболеваний.

Неправильное питание, богатое холестерином, насыщенными жирами и транс-жирами, увеличивает риск сердечно-сосудистых заболеваний.

Отсутствие физической активности может быть самостоятельным фактором риска для сердечных заболеваний и увеличивает вероятность других факторов риска, таких как избыточный вес, гипертония, высокий уровень холестерина и сахарный диабет.

Высокий уровень глюкозы в крови может повреждать кровеносные сосуды и увеличивать риск сердечно-сосудистых заболеваний.

Комбинированный набор этих факторов, называемый метаболическим синдромом, повышает вероятность заболеваний сердца и сосудов.

Стресс является серьезным фактором риска, так как неконтролируемый стресс может привести к неблагоприятным привычкам, таким как курение, употребление алкоголя и переедание, что также увеличивает риск сердечно-сосудистых заболеваний. Стресс также может вызвать повышение артериального давления и даже синдром Такоцубо, известный как "синдром разбитого сердца". Этот синдром характеризуется симптомами, похожими на сердечный приступ, но не связанными с блокадой артерий. Стрессовая кардиомиопатия чаще встречается у женщин.

Не модифицируемые факторы риска сердечно-сосудистых заболеваний включают следующее:

С возрастом увеличивается риск сердечных заболеваний. У мужчин этот риск возрастает после 45 лет, а у женщин - после 55 лет, особенно после наступления менопаузы.

Семейная история играет важную роль. Если близкие родственники (отец, мать, брат, сестра) страдали от сердечных заболеваний, их диагноз был поставлен до 55 лет у мужчин и до 65 лет у женщин, то риск таких заболеваний у вас повышен. В этом случае важно проконсультироваться с врачом, чтобы узнать, когда начинать регулярные обследования и какие факторы риска следует контролировать.

Преэклампсия, осложнение, которое чаще возникает после 20 недели беременности, увеличивает риск сердечно-сосудистых заболеваний в будущем. Преэклампсия характеризуется высоким артериальным давлением, протеином в моче и отечностью. Если не лечить, она может вызвать серьезные осложнения для матери и ребенка.

Заболевание, вызванное новой коронавирусной инфекцией (COVID-19), увеличивает риск тромбозов, которые могут привести к инфаркту миокарда, инсульту и другим осложнениям. Этот риск может возникнуть даже у людей без предыдущих сердечных заболеваний.

Другие факторы риска включают социально-демографические аспекты, такие как низкий уровень образования и дохода, а также проживание в бедных районах,

что увеличивает вероятность развития ишемической болезни сердца. Сильные эмоции, такие как гнев и горе, могут также повысить риск инфаркта миокарда, особенно у людей с существующими сердечно-сосудистыми заболеваниями. Хронический стресс на работе, такой как переработка и длительные рабочие дни, может способствовать развитию ишемической болезни сердца у мужчин. Долгосрочные стрессовые ситуации в семье также могут увеличить этот риск. Клиническая депрессия, панические атаки и тревожность также входят в число факторов риска.

Таким образом, риск развития сердечного приступа является наиважнейшей частью современного медицинского мира. Он не мало опасен для человека и может быть подвергнут корреляционно-регрессионному анализу.

1.2 Корреляционно-регрессионный анализ

Корреляционно-регрессионный анализ является наиболее широко распространенным и гибким приемом обработки статистической информации.

Корреляционно-регрессионный анализ — это один из самых распространенных методов изучения отношений между численными величинами. Его основная цель состоит в нахождении зависимости между двумя параметрами и ее степени с последующим выведением уравнения. То есть, корреляционно-регрессионный анализ представляет из себя объединение методов корреляционного и регрессионного анализов.

Задачами корреляционно-регрессионного анализа являются:

1. Установление типа уравнения регрессии;
2. Определение параметров уравнения регрессии и оценка значимости параметров;
3. Оценка тесноты и направления связи между переменными; - оценка значимости уравнения регрессии;
4. Определение прогнозных значений зависимой переменной и оценка полученного прогноза.

Так как в корреляционно-регрессионном анализе используются методы корреляционного и регрессионного анализа, рассмотрим эти методы подробней.

Корреляционный анализ — раздел математической статистики, в котором изучаются задачи выявления статистических зависимостей между случайными величинами путем оценок различных коэффициентов корреляции. Методы корреляционного анализа дают хорошие результаты тогда, когда данные эксперимента можно считать выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

Невозможно управлять явлениями, предсказывать их развитие без изучения характера, силы и других особенностей связей. Поэтому методы исследования, изменения связей составляют чрезвычайно важную часть методологии научного исследования, в том числе и статистическую.

Связи между изучаемыми переменными подразделяются на функциональные и статистические. При функциональной связи определенному значению одной переменной величины соответствует строго определенное значение другой переменной.

При изменении одной из них на определенную величину, другая переменная изменяется на величину, в соответствии с видом функции, связывающей переменные.

Статистической называется связь между переменными или признаками, когда определенному значению факторного признака соответствует несколько различных значений результативного признака. Частным случаем статистической связи является корреляционная, которая проявляется в среднем, в массе наблюдений, как статистическая закономерность.

При корреляционной связи с изменением факторного признака на определенную величину изменяется среднее значение результативного признака. Обычно корреляционная зависимость представляется как функциональная зависимость между переменными в виде уравнения регрессии.

Корреляционной связью называют важнейший частный случай статистической связи, состоящий в том, что разным значениям одной переменной соответствуют различные средние значения другой. С изменением значения признака x закономерным образом изменяется среднее значение признака y; в то время как в каждом отдельном случае значение признака y (с различными вероятностями) может принимать множество различных значений.

Тесноту связи изучаемых явлений оценивает Коэффициент Пирсона (K_n).

Коэффициент Пирсона используется для изучения связи между двумя качественными признаками, каждый из которых состоит более чем из двух групп. Вычисляют по формуле:

$$K_n = \sqrt{\frac{\varphi^2}{1+\varphi^2}},$$

где φ^2 — показатель взаимной сопряженности:

$$\varphi^2 = \sum \frac{n_{xy}^2}{n_x * n_y},$$

где n_x — объемы признака X по группам;

n_y — объемы признака Y по группам;

n_{xy} — объемы выборок, относящихся к X и Y одновременно.

Корреляционный коэффициент Пирсона может принимать значения в диапазоне $-1 < K_n < 1$.

По значению эмпирического корреляционного отношения судят о тесноте связи между признаками. Обычно придерживаются следующей шкалы:

$0{,}1 < K_n \leq 0{,}3$ — связь слабая;

$0{,}3 < K_n \leq 0{,}5$ — связь заметная;

$0{,}5 < K_n \leq 0{,}7$ — связь умеренно тесная;

$0{,}7 < K_n \leq 0{,}9$ — связь тесная;

$K_n > 0{,}9$ — связь очень тесная.

После того как с помощью корреляционного анализа выявлено наличие статистических связей между переменными и оценена степень их тесноты, обычно переходят к математическому описанию зависимостей, то есть к регрессионному анализу.

Регрессионный анализ применяется в тех случаях, когда необходимо отыскать непосредственно вид зависимости x и y. При этом предполагается, что независимые факторы являются не случайными величинами, а результативный показатель y имеет постоянную, независимую от факторов дисперсию и стандартное отклонение.

Рассмотрим метод линейной регрессии.

Под линейностью имеется в виду, что переменная y предположительно находится под влиянием переменной x в зависимости

$$y_i = a * x_i + b + \varepsilon_i,$$

где b — постоянная величина (или свободный член уравнения);

a — коэффициент регрессии, определяющий наклон линии, вдоль которой рассеяны данные наблюдения. Это показатель, характеризующий изменение переменной y_i при изменении значения x_i на единицу. Если $a > 0$, переменные x_i и y_i положительно коррелированные, если $a < 0$ — отрицательно коррелированные;

ε_i — независимая нормально распределенная величина — остаток с нулевым математическим ожиданием и постоянной дисперсией:

$$D_B = \sigma^2,$$

где σ — среднее квадратическое отклонение. Отражает тот факт, что изменение y_i будет неточно описываться изменением x_i: присутствуют другие факты, не учтенные в данной модели.

Построение линейной регрессии сводится к оценке параметров a и b. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных (теоретических) $\widehat{y_i}$, минимальна:

$$\sum(y_i - \widehat{y_i})^2 \to min,$$

где y_i — фактические значения результативного признака y;

$\widehat{y_i}$ — расчетные значения результативного признака y.

То есть из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной.

Таким образом, рассмотрено понятие корреляционно-регрессионного анализа, а также методы корреляционно-регрессионного анализа: нахождение корреляционной связи с помощью коэффициента Пирсона; построение парной линейной регрессии с помощью метода наименьших квадратов (МНК).

Рассмотрим применение корреляционно-регрессионного анализа на примере анализа риска развития сердечного приступа.

2 Практическая часть

2.1 Корреляционно-регрессионный анализ на примере на примере анализа риска развития сердечного приступа.

2.1.1 О наборе данных

2.1.1.1 Контекст:

Набор данных для прогнозирования риска сердечного приступа служит ценным ресурсом для изучения сложной динамики здоровья сердца и его предикторов. Сердечные приступы, или инфаркты миокарда, по-прежнему остаются серьезной проблемой глобального здравоохранения, требующей более глубокого понимания их предвестников и потенциальных смягчающих факторов. Этот набор данных включает в себя широкий спектр атрибутов, включая возраст, уровень холестерина, кровяное давление, привычки к курению, режим физических упражнений, диетические предпочтения и многое другое, с целью прояснения сложного взаимодействия этих переменных при определении вероятности сердечного приступа. Используя прогностическую аналитику и машинное обучение на основе этого набора данных, исследователи и медицинские работники могут разрабатывать про активные стратегии профилактики и лечения сердечно-сосудистых заболеваний. Этот набор данных является свидетельством коллективных усилий по улучшению нашего понимания здоровья сердечно-сосудистой системы и прокладыванию пути к более здоровому будущему.

2.1.1.2 Содержание:

Этот набор данных предоставляет полный набор характеристик, имеющих отношение к здоровью сердца и выбору образа жизни, включая сведения о конкретном пациенте, такие как возраст, пол, уровень холестерина, кровяное давление, частота сердечных сокращений, а также такие показатели, как диабет, семейный анамнез, привычки к курению, ожирение и потребление алкоголя. Кроме того, учитываются такие факторы образа жизни, как продолжительность физических упражнений, пищевые привычки, уровень стресса и малоподвижный образ жизни. Рассматриваются медицинские аспекты, включающие предшествующие проблемы

с сердцем, прием лекарств и уровень триглицеридов. Учитываются социально-экономические аспекты, такие как доход, и географические атрибуты, такие как страна, континент и полушарие. Набор данных, состоящий из 8763 записей пациентов со всего мира, завершается важнейшим бинарным классификационным признаком, обозначающим наличие или отсутствие риска сердечного приступа, предоставляя всеобъемлющий ресурс для прогностического анализа и исследований в области сердечно-сосудистого здоровья.

2.1.2 Глоссарий набора данных (по столбцам)

1. Patient ID - Уникальный идентификатор для каждого пациента

2. Age - Возраст пациента

3. Sex - Пол пациента (мужчина/женщина)

4. Cholesterol - Уровень холестерина у пациента

5. Blood Pressure - Артериальное давление пациента (систолическое/диастолическое)

6. Heart Rate - Частота сердечных сокращений пациента

7. Diabetes - Есть ли у пациента сахарный диабет (Да/Нет)

8. Family History - Семейный анамнез проблем, связанных с сердцем (1: Да, 0: Нет)

9. Smoking - Статус курения пациента (1: Курильщик, 0: Некурящий)

10. Obesity - Статус ожирения пациента (1: Ожирение, 0: Отсутствие ожирения)

11. Alcohol Consumption - Уровень потребления алкоголя пациентом (Отсутствует/Легкий/УмеренныйТяжелый)

12. Exercise Hours Per Week - Количество часов занятий спортом в неделю

13. Diet - Пищевые привычки пациента (Здоровые/Средние/Нездоровые)

14. Previous Heart Problems - Предыдущие проблемы с сердцем у пациента (1: Да, 0: Нет)

15. Medication Use - Прием лекарств пациентом (1: Да, 0: Нет)

16. Stress Level - Уровень стресса, о котором сообщил пациент (1-10)

17. Sedentary Hours Per Day - Часы сидячей деятельности в день

18. Income - Уровень дохода пациента

19. BMI - Индекс массы тела (ИМТ) пациента

20. Triglycerides - Уровень триглицеридов у пациента

21. Physical Activity Days Per Week - Дни физической активности в неделю

22. Sleep Hours Per Day - Количество часов сна в сутки

23. Country - Страна пациента

24. Continent - Континент, на котором проживает пациент

25. Hemisphere - Полушарие, в котором находится пациент

26. Heart Attack Risk - Наличие риска сердечного приступа (1: Да, 0: Нет)

2.1.3 Предварительный анализ и подготовка данных

Для обработки данных и проведения корреляционно-регрессионного анализа используем среду разработки Jupyter Notebook и язык программирования Python с его библиотеками для анализа данных. Импортируем наши наборы данных в формате CSV. Для этого используем библиотеку Pandas.

Посмотрим на табличное представление исходных данных на рисунке 2.1.

Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	Alcohol Consumption	Exercise Hours Per Week	Diet	Previous Heart Problems	Medication Use	Stress Level	Sedentary Hours Per Day	Income	BMI	Triglycerides	Physical Activity Days Per Week	Sleep Hours Per Day	Country	Continent	Hemisphere	Heart Attack Risk
HQM9364	75	Female	136	141/85	101	No	No	0	1	Light	14.744881	Unhealthy	0	1	4	10.922177	94152	30.589796	374	3	4	Nigeria	Africa	Northern Hemisphere	1
ESJ9954	62	Male	262	137/82	89	Yes	No	0	1	Light	16.228489	Unhealthy	0	1	7	1.208610	159792	31.584511	678	0	8	Australia	Australia	Southern Hemisphere	1
ONA1218	72	Female	126	138/93	86	Yes	Yes	1	0	None	6.818887	Average	0	1	4	9.514556	254952	34.711478	736	1	5	Argentina	South America	Southern Hemisphere	1
UBE5339	18	Female	300	132/94	109	Yes	No	1	1	Light	18.297860	Average	1	1	6	9.015221	25229	29.022289	152	2	5	Spain	Europe	Southern Hemisphere	1
LUQ7367	67	Female	223	91/89	84	Yes	Yes	1	0	Moderate	10.980701	Average	0	1	4	10.020410	229179	35.966244	744	5	8	Japan	Asia	Northern Hemisphere	1

Рисунок 2.1 – Табличное представление исходных данных.

Чтобы получше узнать какого рода данные представлены в таблице посмотрим на их тип.

На рисунке 2.2 указаны типы данных для каждой переменной.

```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Patient ID                      8763 non-null   object
 1   Age                             8763 non-null   int64
 2   Sex                             8763 non-null   int64
 3   Cholesterol                     8763 non-null   int64
 4   Blood Pressure                  8763 non-null   object
 5   Heart Rate                      8763 non-null   int64
 6   Diabetes                        8763 non-null   int64
 7   Family History                  8763 non-null   int64
 8   Smoking                         8763 non-null   int64
 9   Obesity                         8763 non-null   int64
 10  Alcohol Consumption             8763 non-null   int64
 11  Exercise Hours Per Week         8763 non-null   float64
 12  Diet                            8763 non-null   object
 13  Previous Heart Problems         8763 non-null   int64
 14  Medication Use                  8763 non-null   int64
 15  Stress Level                    8763 non-null   int64
 16  Sedentary Hours Per Day         8763 non-null   float64
 17  Income                          8763 non-null   int64
 18  BMI                             8763 non-null   float64
 19  Triglycerides                   8763 non-null   int64
 20  Physical Activity Days Per Week 8763 non-null   int64
 21  Sleep Hours Per Day             8763 non-null   int64
 22  Country                         8763 non-null   object
 23  Continent                       8763 non-null   object
 24  Hemisphere                      8763 non-null   object
 25  Heart Attack Risk               8763 non-null   int64
dtypes: float64(3), int64(17), object(6)
```

Рисунок 2.2 – Типы данных для каждой перменной.

В нашей ситуации мы можем полностью избавиться от номера пациента, т.к. для нас это не имеет значения, так же можно опустить переменные страны, полушария и континента. Из-за того, что категориальные признаки сложнее поддаются регрессионному анализу, чем числовые мы можем заменить, где это возможно на числа, таким образом мы заменили в нашей таблице:

1. Значение переменной “Sex”, где значение 1 – “Male”, 0 – “Female”.
2. Значение переменной “Blood Pressure” – разделили по символу “/” на две переменных, а именно “Systolic BP” и “Diastolic BP”.
3. Значение переменной “Diet”, где значение “Unhealthy” – -1, “Average” – 0, “Healthy” – 1.
4. Добавим новую переменную $Chol\_BMI\_Ratio = Cholesterol / BMI$.
5. Значение переменной “Alcohol Consumption”, где значение None – 0, Light – 1, Moderate – 2, Heavy – 3.
6. Значение переменной “Diabetes”, где значение “Yes” – 1, “No” – 0.

Так как эти данные были взяты из открытых источников, в них могут быть пропущены значения, а также на местах значений стоять нули, нам необходимо проверить наличие некачественных данных и задать способ их обработки – например для строк, в которых пропущено какое-то значение можно, либо попробовать восстановить это значение, либо удалить всю строчку.

Проведя первичную очистку данных от некачественных значений и пропусков, а также замены категориальных признаков на числовые получим “чистую” таблицу исходных данных, представленную на рисунке 2.3.

	Age	Sex	Cholesterol	Heart Rate	Diabetes	Family History	Smoking	Obesity	Alcohol Consumption	Exercise Hours Per Week	Diet	Previous Heart Problems	Medication Use	Stress Level	Sedentary Hours Per Day	Income	
0	67	1	208	72	0	0	1	0	0	4.168189	0	0	0	9	6.615001	261404	31
1	21	1	389	98	1	1	1	1	1	1.813242	-1	1	0	1	4.963459	285768	27
2	21	0	324	72	1	0	0	0	0	2.078353	1	1	1	9	9.463426	235282	28
3	84	1	383	73	1	1	1	0	1	9.828130	0	1	0	9	7.648981	125640	36
4	66	1	318	93	1	1	1	1	0	5.804299	-1	1	0	6	1.514821	160555	21
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
8758	60	1	121	61	1	1	1	0	1	7.917342	1	1	1	8	10.806373	235420	19
8759	28	0	120	73	1	0	0	1	0	16.558426	1	0	0	8	3.833038	217881	23
8760	47	1	250	105	0	1	1	1	1	3.148438	0	1	0	5	2.375214	36998	35
8761	36	1	178	60	1	0	1	0	0	3.789950	-1	1	1	5	0.029104	209943	27
8762	25	0	356	75	1	1	0	0	1	18.081748	1	0	0	8	9.005234	247338	32

Рисунок 2.3 – Очищенная таблица исходных данных.

Перед непосредственным анализом, давайте посмотри на общие характеристики наших данных. На количество строк в каждом столбце, среднее значение, медиану, максимальный и минимальный элемент, а также 25-ый, 50-ый и 75-ый процентили представлены на рисунке 2.4.

]:

	Age	Sex	Cholesterol	Heart Rate	Diabetes	Family History	Smoking	Obesity	Alcohol Consumption	Exercise Hours Per Week	Previous Heart Problems
count	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000
mean	53.707977	0.697364	259.877211	75.021682	0.652288	0.492982	0.896839	0.501426	0.598083	10.014284	0.495835
std	21.249509	0.459425	80.863276	20.550948	0.476271	0.499979	0.304186	0.500026	0.490313	5.783745	0.500011
min	18.000000	0.000000	120.000000	40.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.002442	0.000000
25%	35.000000	0.000000	192.000000	57.000000	0.000000	0.000000	1.000000	0.000000	0.000000	4.981579	0.000000
50%	54.000000	1.000000	259.000000	75.000000	1.000000	0.000000	1.000000	1.000000	1.000000	10.069559	0.000000
75%	72.000000	1.000000	330.000000	93.000000	1.000000	1.000000	1.000000	1.000000	1.000000	15.050018	1.000000
max	90.000000	1.000000	400.000000	110.000000	1.000000	1.000000	1.000000	1.000000	1.000000	19.998709	1.000000

Рисунок 2.4 – Общие характеристики данных.

2.1.4 Демографический анализ сердечного приступа

Демографический анализ — это изучение характеристик населения, таких как возраст, пол, раса, этническая принадлежность, доход и уровень образования. Он может быть использован для выявления тенденций и закономерностей в популяции и для понимания того, как эти факторы могут влиять на различные исходы, такие как риск сердечного приступа.

2.1.4.1 Возраст в зависимости риска сердечного приступа

Построим гистограмму распределения количества сердечных приступов от возраста и два “ящика с усами” – один для тех, у кого был сердечный приступ и для тех, у кого его не было представлено на рисунке 2.5.

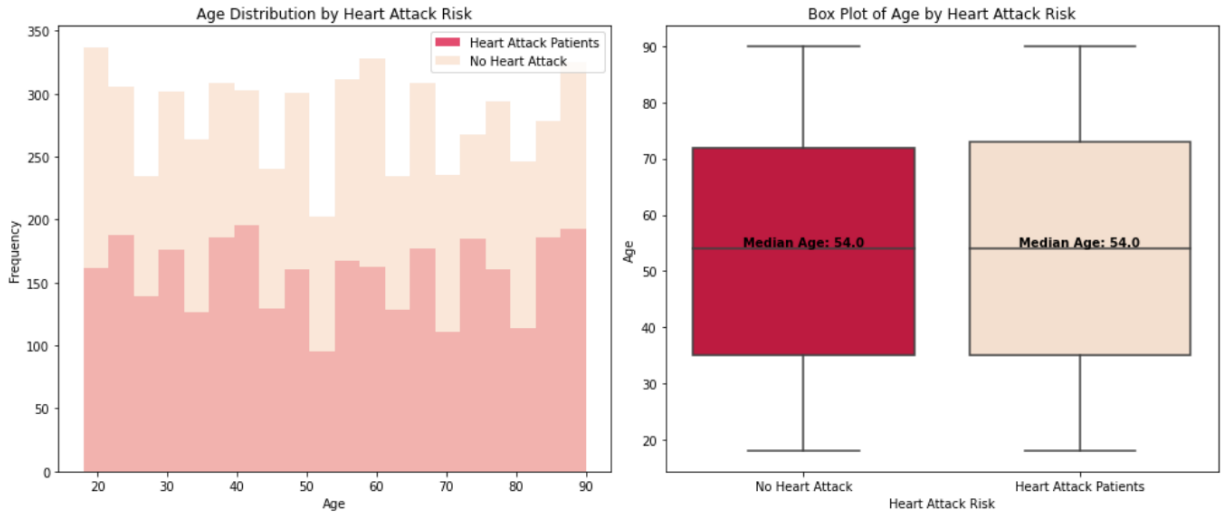

Рисунок 2.5 Графики зависимости возраста и сердечного приступа.

Тем самым по гистограмме можно заметить относительно равное распределение между риском возникновения сердечного приступа для каждого возраста, а также по графикам “boxplot” узнаем медиану среди людей, у которых был сердечный приступ и нет, она оказалась равной в обоих случаях числу 54.

2.1.4.2 Половое распределение

Построим гистограмму распределения в зависимости от пола пациента и риска возникновения сердечного приступа и круговую диаграмму отношения мужчин и

женщин с риском возникновения инфаркта. Диаграммы представлены на рисунках 2.6 и 2.7 соответственно.

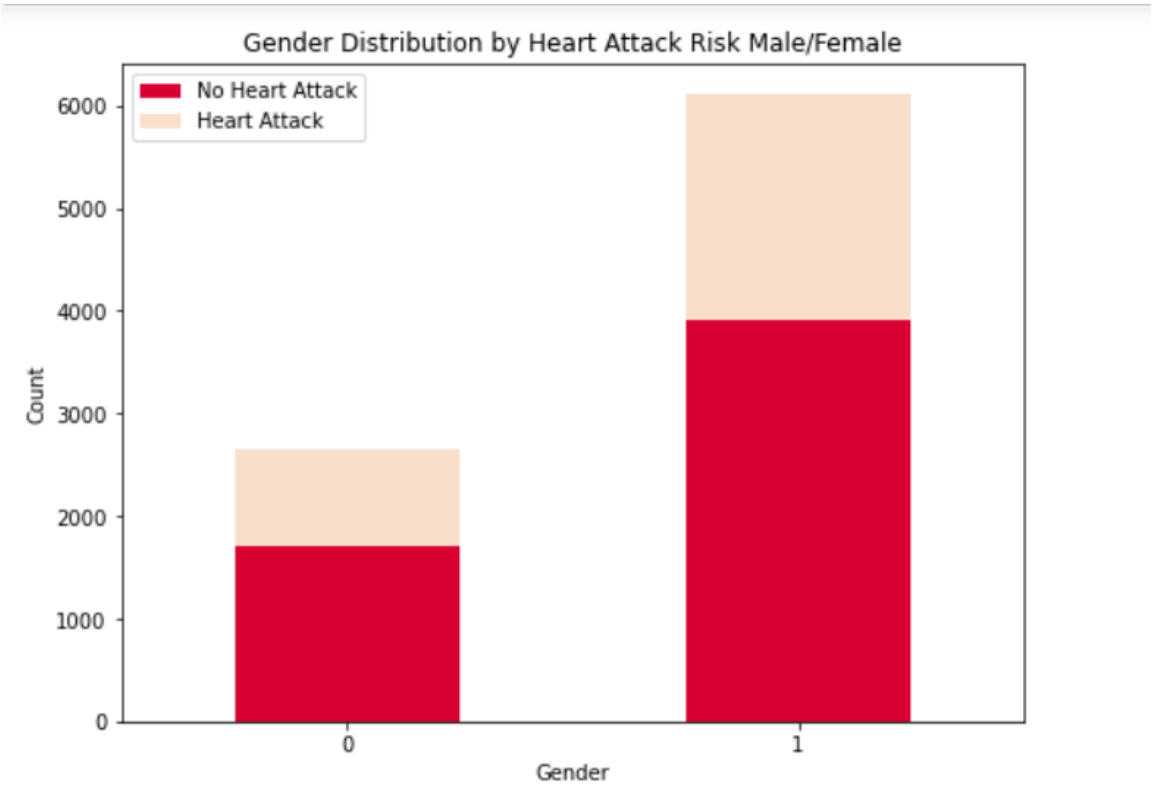

Рисунок 2.6 - Гистограмма распределения зависимости сердечного приступа от пола.

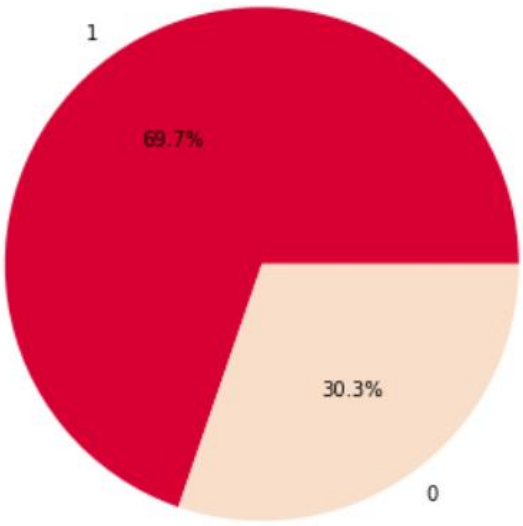

Рисунок 2.7 - Круговая диаграмма отношения мужчин и женщин, имеющих риск к возникновению сердечного приступа.

По представленным графикам заметим, что мужчины чаще подвержены риску возникновения инфаркта, чем женщины. Давайте исследуем биологические факторы для того, чтобы углубиться в детали, возможно узнаем почему мужчины чаще страдают от сердечных приступов.

2.1.5 Биометрические факторы

Исследование биометрических факторов является одним из самых важных моментов при прогнозировании сердечных приступов.

2.1.5.1 Уровень холестерина

Построим два “ящика с усами” зависимости уровня холестерина и сердечного приступа, а также график плотности распределения холестерина. Данные графики представлены на рисунках 2.8 и 2.9 соответственно.

Рисунок 2.8 – График зависимости уровня холестерина и сердечного приступа.

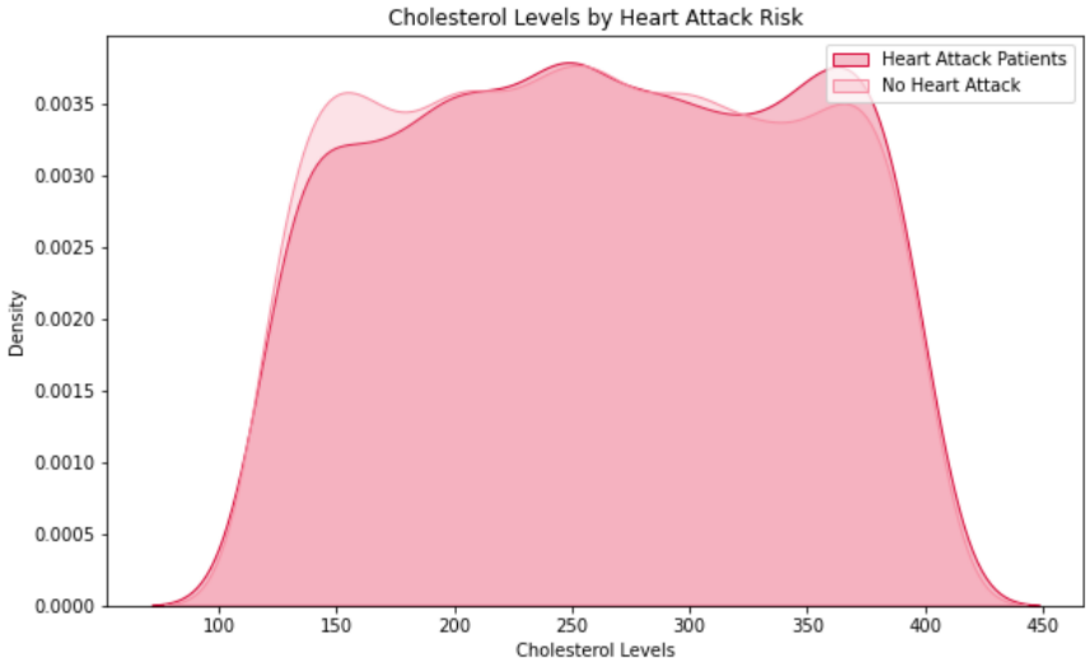

Рисунок 2.9 – График плотности распределения холестерина.

По “ящикам с усами” видно небольшое смещение правого “ящика” выше относительного левого, тем самым можно понять, что его значения чуть выше, что подтверждается на графике плотности, тем самым можно предположить, что чем выше уровень холестерина, тем выше риск возникновения инфаркта.

2.1.5.2 Кровеносное давление

Построим два графика для систолического давления и диастолического давления отдельно по два ящика с усами на каждый разделив их по принадлежности к полу на рисунках 2.10 и 2.11 соответственно.

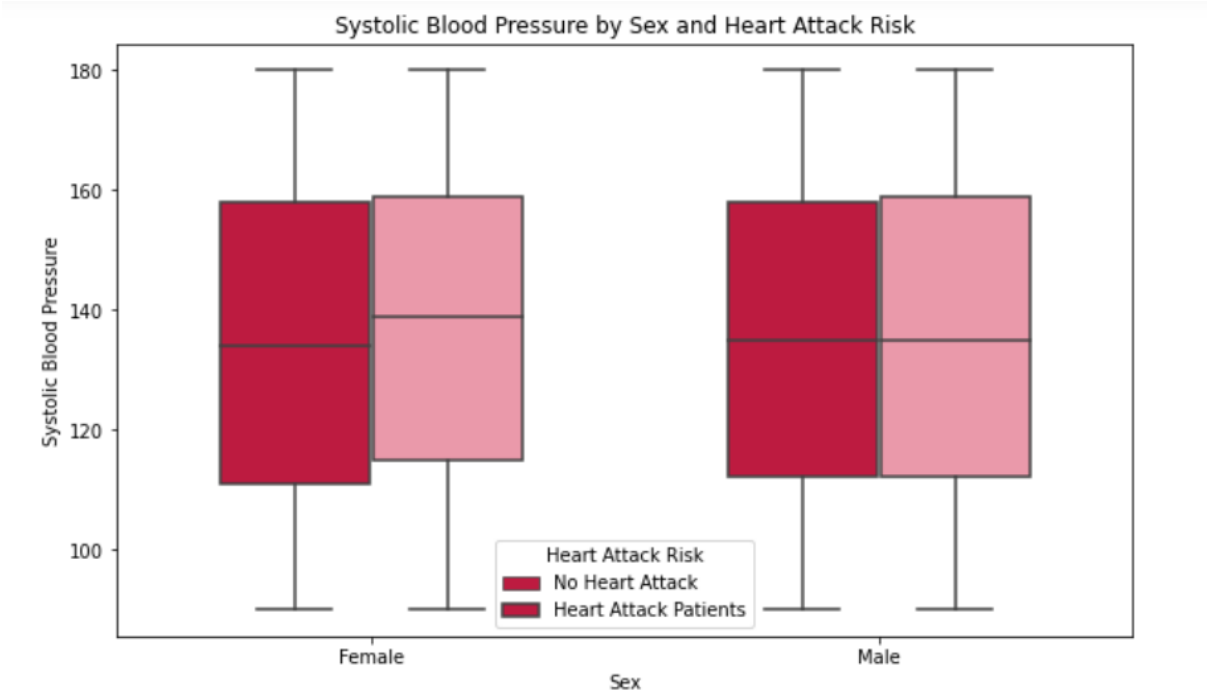

Рисунок 2.10 – Систолическое давление по полу.

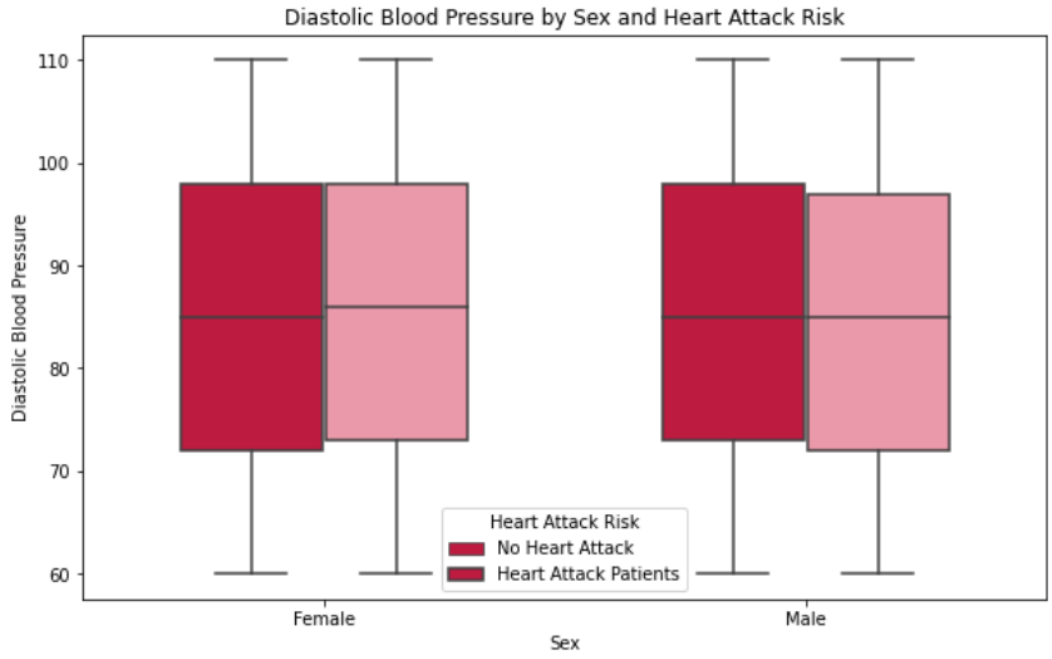

Рисунок 2.11 – Диастолическое давление по полу.

По графикам делаем вывод, что среди женщин чаще подвержены риску возникновения инфаркта те, у кого систолическое давление ниже, а среди мужчин

больше подвержены риску сердечного приступа, чье диастолическое давление выше.

2.1.5.3 Частота сердцебиения

Построим гистограмму зависимости частоты распространения риска инфаркта по частоте сердцебиения представленную на рисунке 2.12.

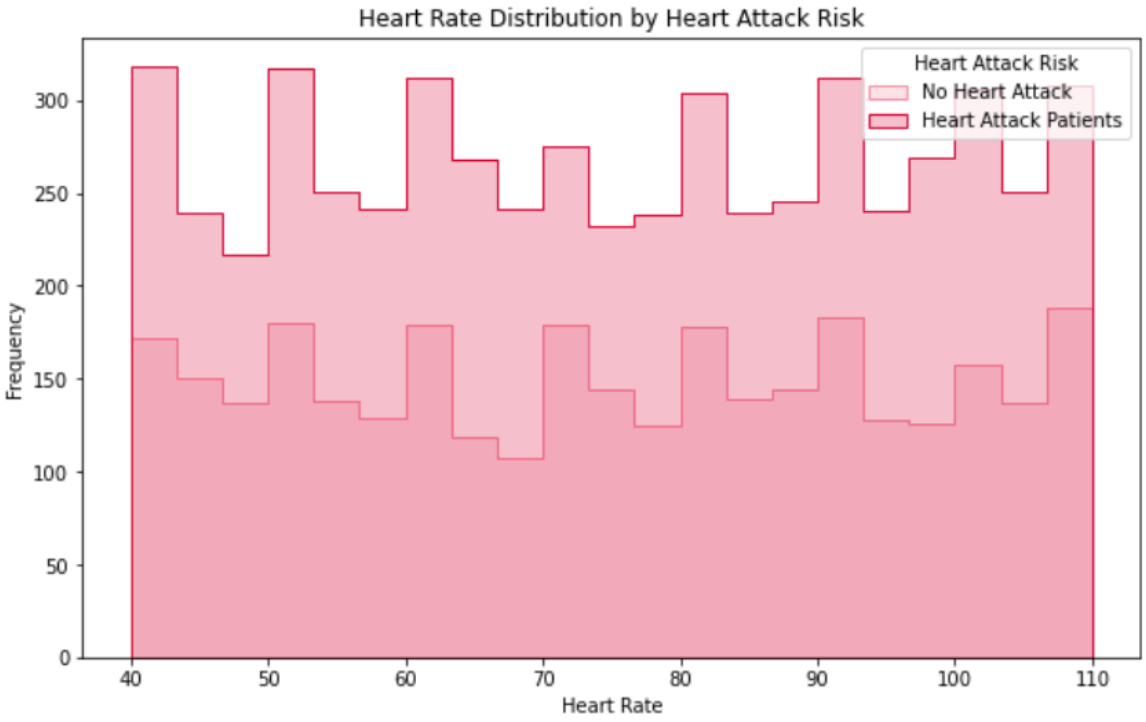

Рисунок 2.12 – Частота распространения риска инфаркта по частоте сердцебиения.

2.1.6 Образ жизни и поведение в отношение здоровья

2.1.6.1 Влияние курения, ожирения и алкоголя

Построим гистограммы соотношения курения, ожирения и алкоголя к риску сердечных приступов, представленных на рисунке 2.13.

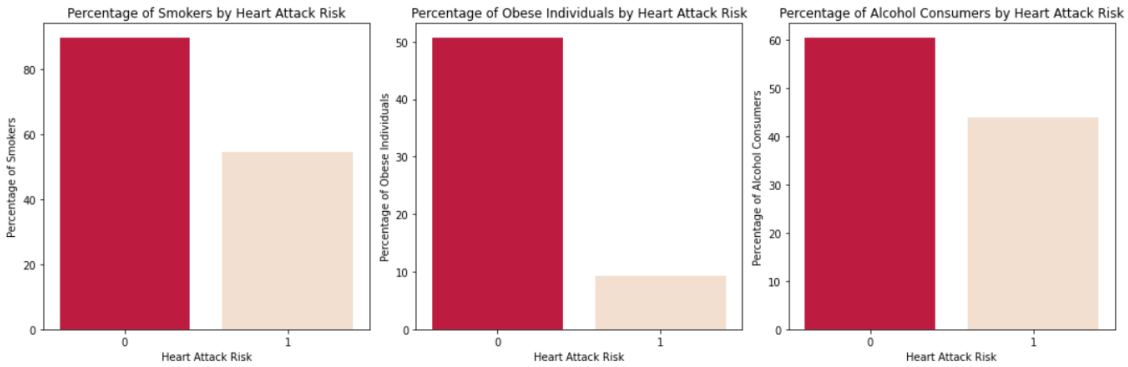

Рисунок 2.13 – Зависимости влияния курения, ожирения и алкоголя на риск возникновения сердечного приступа.

С легкостью можно заметить, что если человек имеет за собой вредные привычки или проблемы с лишнем весом, он больше подвержен риску сердечного приступа.

2.1.7 История болезней и состояния

2.1.7.1 Диабет, семейный анамнез и предыдущие проблемы с сердцем

Построим гистограммы суммы пропорций людей с подверженным сердечному приступу по отношению к наличию диабета, плохому семейному анамнезу и наличием в прошлом проблем с сердцем, гистограмма представлена на рисунке 2.14.

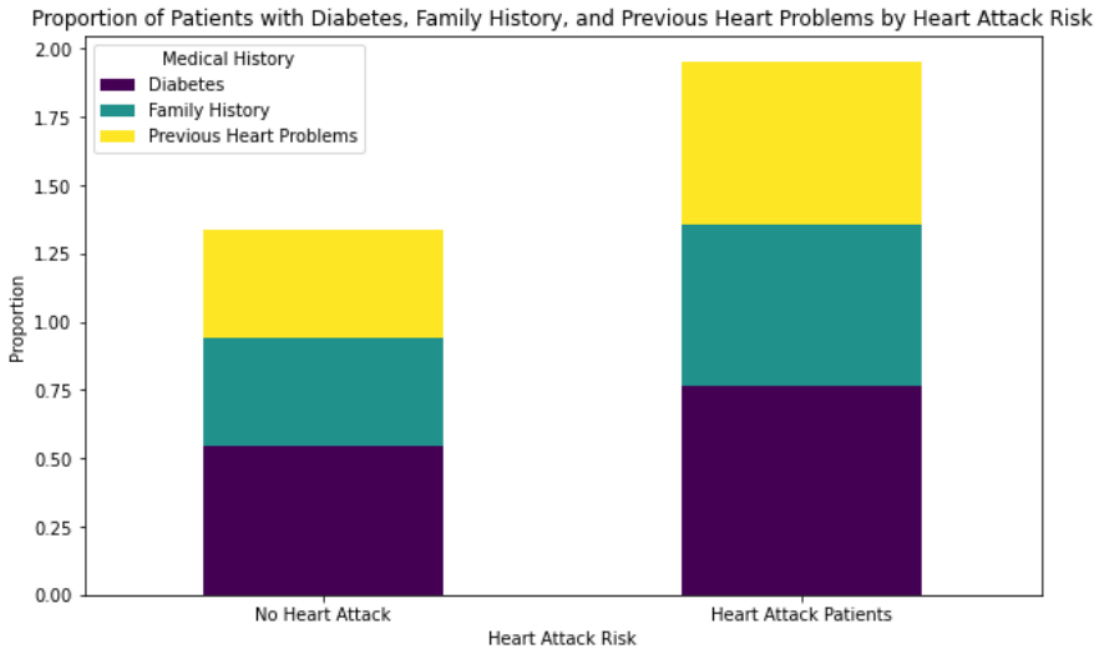

Рисунок 2.14 – Гистограмма суммы пропорций состояния пациента.

2.1.7.2 Прием лекарств и триглицериды

Построим гистограмму приема лекарств и плотности распределения триглицеридов на рисунке 2.15.

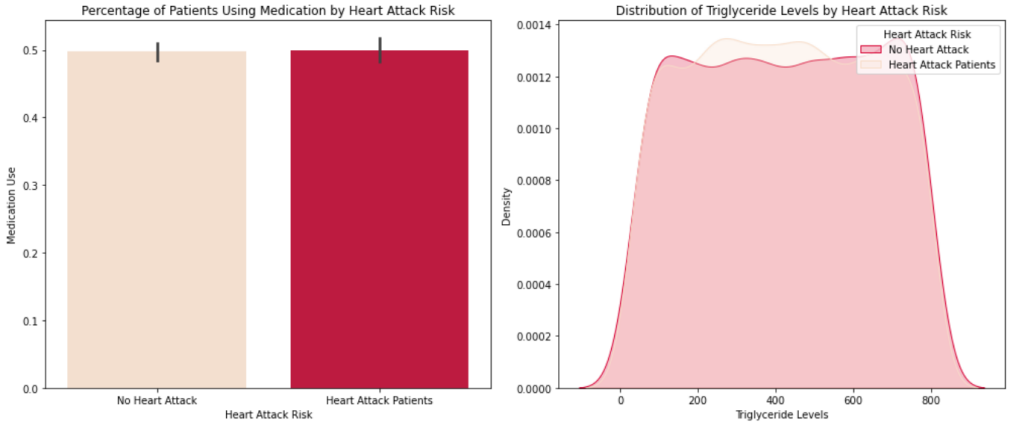

Рисунок 2.15 – Графики приема лекарств и триглицеридов.

2.1.8 Корреляции данных и тренды

2.1.8.1 Корреляционная тепловая карта

Построим тепловую корреляционную карту и представим на рисунке 2.16.

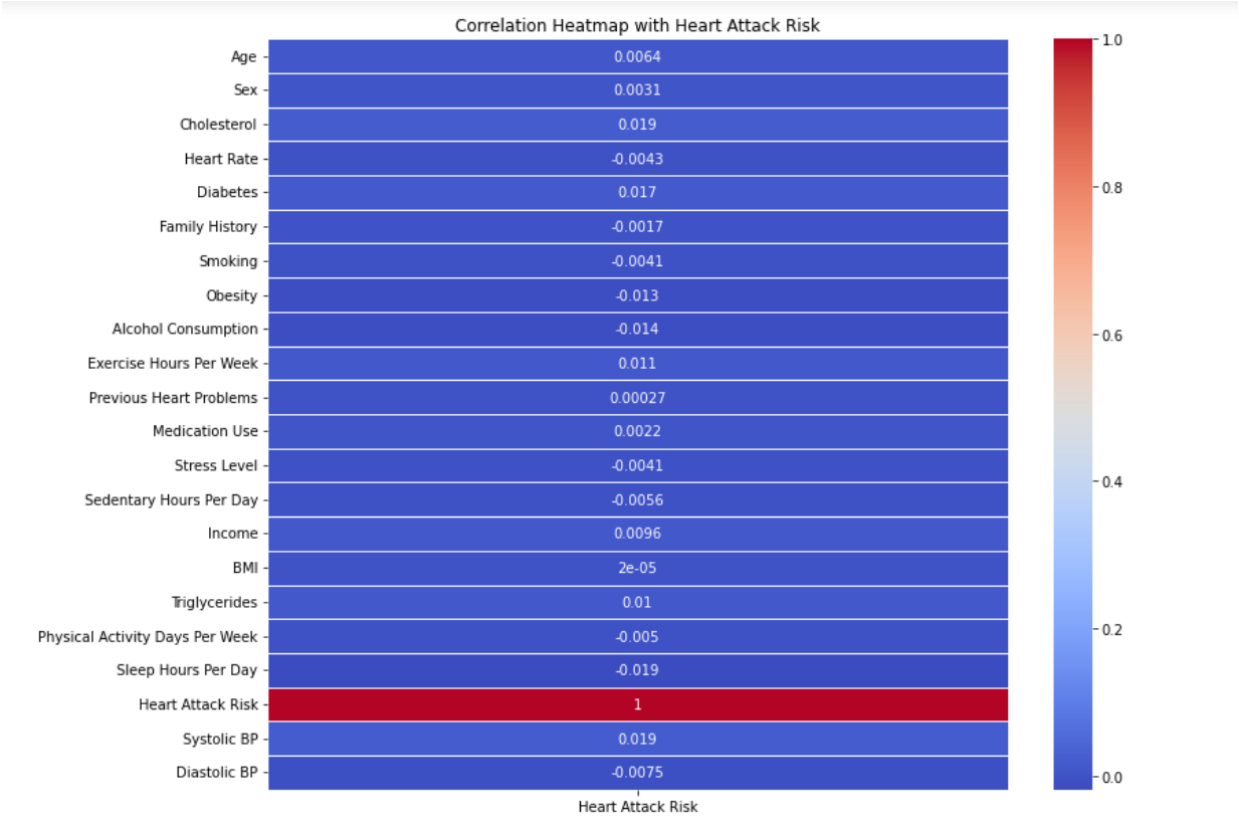

Рисунок 2.16 – Корреляционная тепловая карта.

По ней можно заметить те, значения, которые лучше всего коррелируют с нашим целевым признаком.

2.1.8.2 Тенденции с течением времени

Построим график зависимости среднего значения риска сердечного приступа от возраста, выделив интересующую для нас зону в окрестности ранее найденного значения 54 (Рисунок 2.5) представим график на рисунке 2.17.

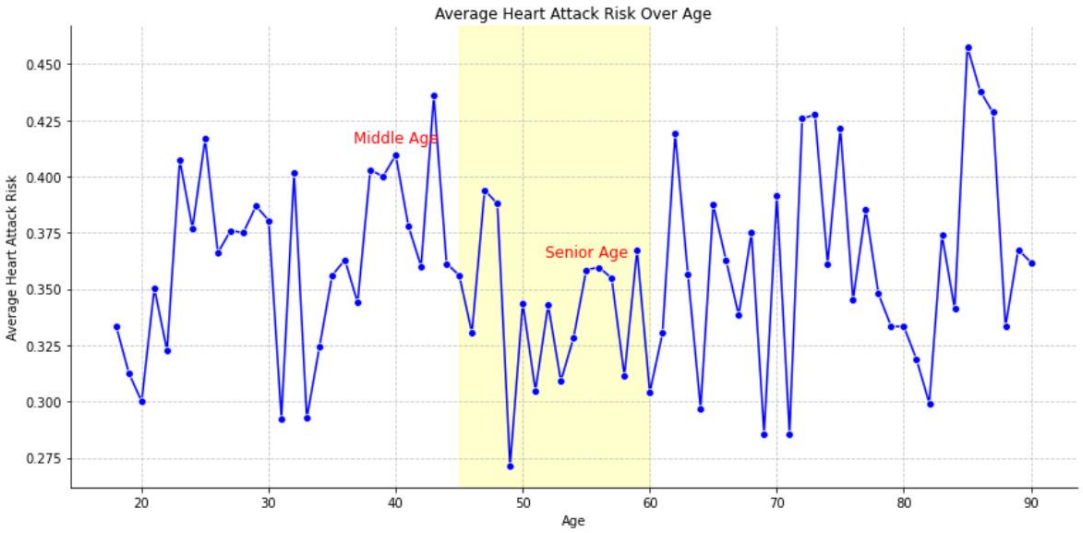

Рисунок 2.17 – Тенденции с течением времени.

Завершая предварительный этап анализа данных, мы уже нашли зависимости целевой переменной от других и на этой основе можем делать заключения полезные для дальнейшей предиктивной аналитике.

2.2 Предиктивная аналитика

Для того чтобы предсказывать есть ли у пациента риск возникновения инфаркта необходимо построить модель машинного обучения, но прежде необходимо нормализовать некоторые значения. Ключевая цель нормализации — приведение различных данных в самых разных единицах измерения и диапазонах значений к единому виду, который позволит сравнивать их между собой или использовать для расчёта схожести объектов.

2.2.1 Нормализация данных

Использовать будем StandardScaler — это метод масштабирования на основе среднего. Формула $StandardScaler = (X_i - X_{mean}) \,/\, X_{std}$, поэтому он устанавливает среднее значение как 0.

StandardScaler уязвим для выбросов, поскольку выбросы влияют на среднее значение, но в нашем случае это не страшно, т.к. мы их заранее обработали. Если у вас нормальное распределение или данные, близкие к нормальным, StandardScaler приближает ваши данные к стандартному нормальному распределению.

После нормализации получим таблицу, приведенную на рисунке 2.18.

	Age	Sex	Cholesterol	Heart Rate	Diabetes	Family History	Smoking	Obesity	Alcohol Consumption	Exercise Hours Per Week	Diet	Previous Heart Problems	Medication Use	Stress Level	Sedentary Hours Per Day	I
0	0.625557	1	-0.641579	-0.147042	0	0	1	0	0	-1.010838	0	0	0	9	0.179251	1.2
1	-1.539322	1	1.596895	1.118179	1	1	1	1	1	-1.418027	-1	1	0	1	-0.297225	1.5
2	-1.539322	0	0.793023	-0.147042	1	0	0	0	0	-1.372188	1	1	1	9	1.001031	0.9
3	1.425621	1	1.522691	-0.098380	1	1	1	0	1	-0.032188	0	1	0	9	0.477557	-0.4
4	0.578495	1	0.718820	0.874867	1	1	1	1	0	-0.727941	-1	1	0	6	-1.292170	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
8758	0.296119	1	-1.717530	-0.682328	1	1	1	0	1	-0.362578	1	1	1	8	1.388476	0.9
8759	-1.209884	0	-1.729898	-0.098380	1	0	0	1	0	1.131536	1	0	0	8	-0.623356	0.7
8760	-0.315695	1	-0.122154	1.458815	0	1	1	1	1	-1.187161	0	1	0	5	-1.043943	-1.5
8761	-0.833383	1	-1.012597	-0.730990	1	0	1	0	0	-1.076238	-1	1	1	5	-1.720804	0.6
8762	-1.351072	0	1.188775	-0.001055	1	1	0	0	1	1.394931	1	0	0	8	0.868841	1.1

Рисунок 2.18 – Фрагмент нормализованной таблице данных.

2.2.2 Разделение выборки

Разделим наши данные на две выборки – тренировочную и тестовую в соотношении 80 на 20. Как работает 80/20? В большинстве случаев, около 80% результата происходят из 20% причин. Эти цифры могут меняться – иногда это 70/30 и иногда 90/10. Но какое бы ни было соотношение, суть в том, что небольшое количество причин оказывают непропорционально большое влияние на результат.

Тем самым получим тензоры размерностей, представленные на рисунке 2.19.

```
X_train shape: (5842, 23)
X_test shape: (2921, 23)
y_train shape: (5842,)
y_test shape: (2921,)
```

Рисунок 2.19 – Размерности тензоров.

Подготовительный этап полностью завершен теперь можем приступить к созданию моделей.

2.2.3 Способ первый – Обычная линейная регрессия

Регрессионная модель

$$y = f(x, b) + \varepsilon, \quad E(\varepsilon),$$

где b — параметры модели, ε — случайная ошибка модели; называется линейной регрессией, если функция регрессии $f(x, b)$ имеет вид

$$f(x, b) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k ,$$

где b_j — параметры (коэффициенты) регрессии, x_j — регрессоры (факторы модели), k — количество факторов модели.

Коэффициенты линейной регрессии показывают скорость изменения зависимой переменной по данному фактору, при фиксированных остальных факторах (в линейной модели эта скорость постоянна):

$$\forall j \quad b_j = \frac{\partial f}{\partial x_j} = const,$$

Параметр b_0, при котором нет факторов, называют часто константой. Формально — это значение функции при нулевом значении всех факторов. Для аналитических целей удобно считать, что константа — это параметр при «факторе», равном 1 (или другой произвольной постоянной, поэтому константой называют также и этот «фактор»). В таком случае, если перенумеровать факторы и параметры исходной модели с учетом этого (оставив обозначение общего количества факторов — k), то линейную функцию регрессии можно записать в следующем виде, формально не содержащем константу:

$$f(x, b) = b_1 x_1 + b_2 x_2 + \cdots + b_k x_k = \sum_{j=1}^{k} b_j x_j = x^T b,$$

где $x^T = (x_1, x_2, \ldots, x_k)$ — вектор регрессоров, $b = (b_1, b_2, \ldots, b_k)^T$ — вектор-столбец параметров (коэффициентов).

Линейная модель может быть как с константой, так и без константы. Тогда в этом представлении первый фактор либо равен единице, либо является обычным фактором соответственно.

Построив модель линейной регрессии, получили набор из значений, представленных на рисунке 2.20 и значение средней квадратичной ошибки на рисунке 2.21.

	Actual	Predicted
1226	0	0.3764254289
7903	1	0.4270440674
1559	1	0.3741702110
3621	1	0.3668674338
7552	0	0.3564891165

Рисунок 2.20 – Набор предсказаний.

MSE оценивает качество либо предсказателя (т. е. функции, отображающей произвольные входные данные в выборку значений некоторой случайной величины), либо оценщика (т.е. математической функции, отображающей выборку данных в оценку параметра совокупности, из которой отбираются данные). В контексте прогнозирования понимание интервала прогнозирования также может быть полезным, поскольку оно обеспечивает диапазон, в пределах которого с определенной вероятностью попадет будущее наблюдение. Определение MSE отличается в зависимости от того, описывается ли предиктор или оценщик.

Предсказатель

Если вектор n прогнозов генерируется из выборки n точек данных по всем переменным и Y является вектором наблюдаемых значений прогнозируемой переменной, причем $\hat{Y}$ это предсказанные значения (например, из подбора методом наименьших квадратов), то MSE предсказателя в пределах выборки вычисляется как

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 .$$

Другими словами, MSE является *средним* $(\frac{1}{n}\sum_{i=1}^{n})$ квадратов ошибок $(Y_i - \hat{Y}_i)^2$. Это легко вычисляемая величина для конкретной выборки (и, следовательно, зависит от выборки).

В матричных обозначениях,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2 = \frac{1}{n}\, e^T\, e$$

где e_i - $(Y - \hat{Y})^2$ и e является $n \times 1$ вектором-столбцом.

MSE также может быть вычислен для q точек данных, которые не использовались при оценке модели, либо потому, что они были отложены для этой цели, либо потому, что эти данные были получены недавно. В рамках этого процесса, известного как перекрестная проверка, MSE часто называют тестовой MSE, и вычисляется как

$$MSE = \frac{1}{q}\sum_{i=n+1}^{n+q}(Y_i - \hat{Y}_i)^2$$

Оценщик

MSE оценщика $\hat{\theta}$ относительно неизвестного параметра θ определяется как

$$MSE(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2].$$

Это определение зависит от неизвестного параметра, но MSE является априорным свойством оценщика. MSE может быть функцией неизвестных параметров, и в этом случае любая оценка MSE, основанная на оценках этих параметров, будет функцией данных (и, следовательно, случайной величиной). Если оценка $\hat{\theta}$ получена как выборочная статистика и используется для оценки

некоторого параметра совокупности, то математическое ожидание относится к выборочному распределению выборочной статистики.

MSE можно записать как сумму дисперсии оценщика и квадрата смещения оценщика, предоставляя полезный способ вычисления MSE и подразумевая, что в случае несмещенных оценок MSE и дисперсия эквивалентны.

$$MSE(\hat{\theta}) = Var_{\theta}(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2.$$

Но в реальном случае моделирования MSE можно описать как сложение дисперсии модели, смещения модели и неустранимой неопределенности (см. Компромисс между смещением и дисперсией). Согласно соотношению, MSE оценщиков может быть просто использован для сравнения эффективности, которое включает информацию о дисперсии и смещении оценщика. Это называется критерием MSE.

```
Mean Squared Error: 0.7596584833790148
```

Рисунок 2.21 – значение среднеквадратичной ошибки.

Данные значения нас не устраивают и давайте попробуем улучшить результат более точной моделью.

2.2.4 Способ второй – Метод обратного исключения

Обратное исключение — это метод пошагового выбора признаков, который начинается с полного набора признаков и итеративно удаляет по одному признаку за раз на основе заданного критерия. Математическая концепция, стоящая за обратным исключением, включает подбор модели с использованием всех признаков и удаление признака с наивысшим p-значением (наименьшей значимостью) на каждой итерации.

Алгоритм отбора начинает работу с модели, содержащей все переменные (такая модель называется «полной»). Затем начинает удалять наименее значимые переменные одну за другой до тех пор, пока не будет достигнуто предварительно заданное правило остановки, или пока в модели не останется ни одной переменной. Как и в случае прямого отбора требуется определить наименее значимую переменную на каждом шаге и правило остановки. Метод обратного исключения представлен на рисунке 2.22.

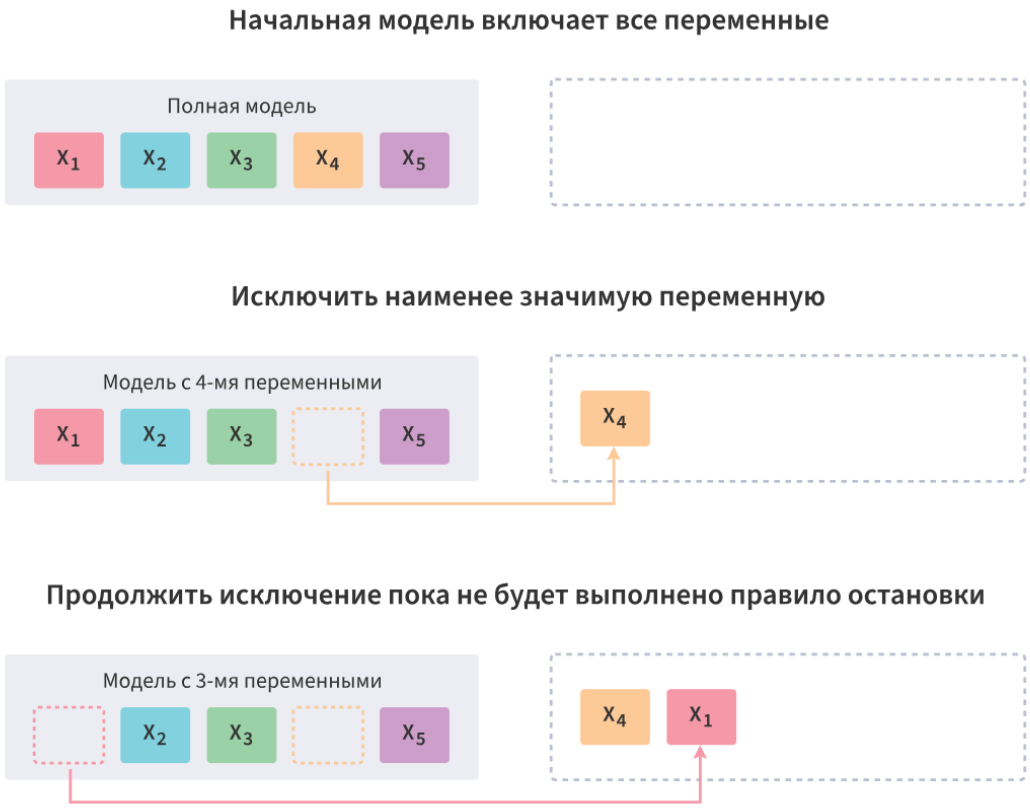

Рисунок 2.22 – Метод обратного исключения.

Очевидно, что первыми кандидатами на исключение являются переменные, которые наименее способствуют повышению качества модели. Аналогично методу прямого включения для оценки значимости изменения качества модели может быть использован критерий Фишера: лучшим кандидатом на исключение будет та переменная, для которой значение критерия Фишера выше заданного порога.

Наименее значимой является переменная:

1. с которой связано наибольшее p-значение;
2. исключение которой из модели вызывает наименьшее сокращение коэффициента детерминации R-квадрат;
3. исключение которой из модели вызывает наименьшее увеличение RSS (суммы квадратов остатков) по сравнению с другими признаками.

Выбор правила остановки

Правило остановки выполняется, когда все оставшиеся переменные в модели имеют p-значение меньше некоторого заранее заданного порога. Когда модель достигнет этого состояния, алгоритм обратного исключения завершится.

Как и в случае прямого выбора, порог может быть:

1. фиксированным значением (например: 0.05, 0.2 или 0.5);
2. определяется AIC;
3. определяется BIC.

Прямой отбор предпочтительно использовать, когда количество рассматриваемых переменных велико. Это связано с тем, что он начинается с нулевой модели и продолжает добавлять переменные по одной, и поэтому, в отличие от обратного отбора, он не рассматривает полную и близкие к ней модели.

Обратный отбор предпочтительно использовать если нужно рассмотреть полную модель, когда одновременном учитываются все переменные. При обратном отборе кандидаты на исключение могут и не появиться и все переменные останутся в модели.

Преимущества пошагового отбора:

1. простота реализации;
2. улучшение интерпретируемости модели;
3. снижение вычислительных затрат за счёт того, что рассматриваются не все переменные;
4. объективность — автоматический выбор позволяет избежать субъективности экспертных оценок.

Особенно оказываются полезными методы пошагового отбора в случае разведочного анализа данных, когда априорные сведения о решаемой задаче отсутствуют.

Недостатки пошагового отбора:

1. не рассматривает все возможные комбинации переменных, поэтому не гарантирует лучшего их набора;
2. приводит к смещенным оценкам коэффициентов регрессии, доверительных интервалов, p-значений и коэффициента R-квадрат;
3. формирует нестабильный набор переменных, особенно в случае, когда число переменных сравнимо с числом наблюдений. Это возможно, когда разные наборы переменных одинаково воздействуют на выходную переменную и выражается в том, что каждый раз получается разный набор переменных. Чтобы избежать данного эффекта требуется, чтобы

число наблюдений выборки на одну входную переменную было 50 и выше.

4. не учитывает причинно-следственные связи между переменными.

Результат второй модели представлен на рисунке 2.23 и 2.24.

Dep. Variable:	y	**R-squared (uncentered):**	0.347
Model:	OLS	**Adj. R-squared (uncentered):**	0.346
Method:	Least Squares	**F-statistic:**	211.5
Date:	Sun, 15 Oct 2023	**Prob (F-statistic):**	0.00
Time:	02:29:05	**Log-Likelihood:**	-6065.7
No. Observations:	8763	**AIC:**	1.218e+04
Df Residuals:	8741	**BIC:**	1.233e+04
Df Model:	22		
Covariance Type:	nonrobust		

Рисунок 2.23 – Результат модели.

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.0091	0.013	0.691	0.489	-0.017	0.035
x2	0.0543	0.020	2.747	0.006	0.016	0.093
x3	-0.0022	0.005	-0.424	0.672	-0.012	0.008
x4	0.0507	0.011	4.796	0.000	0.030	0.071
x5	0.0211	0.010	2.062	0.039	0.001	0.041
x6	0.0964	0.018	5.344	0.000	0.061	0.132
x7	0.0093	0.010	0.908	0.364	-0.011	0.029
x8	0.0134	0.010	1.296	0.195	-0.007	0.034
x9	0.0059	0.005	1.145	0.252	-0.004	0.016
x10	0.0034	0.006	0.533	0.594	-0.009	0.016
x11	0.0227	0.010	2.221	0.026	0.003	0.043
x12	0.0281	0.010	2.758	0.006	0.008	0.048
x13	0.0069	0.002	4.055	0.000	0.004	0.010
x14	-0.0038	0.005	-0.742	0.458	-0.014	0.006
x15	0.0050	0.005	0.969	0.333	-0.005	0.015
x16	-0.0327	0.015	-2.215	0.027	-0.062	-0.004
x17	0.0062	0.005	1.205	0.228	-0.004	0.016
x18	0.0063	0.002	2.845	0.004	0.002	0.011
x19	0.0161	0.002	7.744	0.000	0.012	0.020
x20	0.0092	0.005	1.786	0.074	-0.001	0.019
x21	-0.0041	0.005	-0.793	0.428	-0.014	0.006
x22	-0.0560	0.024	-2.340	0.019	-0.103	-0.009

Рисунок 2.24 – Коэффициенты (веса) регрессии для каждой переменной.

Результат второй модели получился намного лучше по сравнению с первой, удалось сократить квадратичную ошибку в два раза, но это все еще не идеальный результат. Перейдем к третьей заключительной модели.

2.2.5 Третья модель – Технология Catboost от Яндекса

Это открытая программная библиотека, разработанная Яндексом и реализующая уникальный запатентованный алгоритм построения моделей машинного обучения с использованием одной из оригинальных схем повышения градиента.

Воспользуемся моделью градиентного бустинга - это метод машинного обучения, используемый, среди прочего, в задачах регрессии и классификации. Он дает модель прогнозирования в виде ансамбля слабых моделей прогнозирования, то есть моделей, которые делают очень мало предположений относительно данных, которые обычно представляют собой простые деревья решений. Когда дерево решений является слабым учеником, результирующий алгоритм называется деревьями с градиентным усилением; обычно он превосходит случайный лес. Модель деревьев с градиентным усилением строится поэтапно, как и в других методах бустинга, но она обобщает другие методы, позволяя оптимизировать произвольную дифференцируемую функцию потерь.

Как и другие методы бустинга, градиентный бустинг итеративным образом объединяет слабых "учеников" в одного сильного ученика. Это проще всего объяснить в режиме регрессии наименьших квадратов, где цель состоит в том, чтобы "научить" модель F предсказывать значения формы $\hat{y} = F(x)$ путем минимизации среднеквадратичной ошибки $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y_l})^2$, где i индексируются по некоторому обучающему набору величины n фактические значения выходной переменной y:

1. $\widehat{y_l} =$ прогнозируемое значение $F(x_i)$
2. $y_i =$ наблюдаемое значение
3. $n =$ количество выборок в y

Теперь давайте рассмотрим алгоритм градиентного бустинга, состоящий из M этапов. На каждом этапе m $(1 \leq m \leq M)$ повышения градиента предположим, что

некоторая несовершенная модель F_m (для низких m значений эта модель может просто возвращать $\widehat{y_i} = \bar{y}$, где RHS является средним значением y). Для улучшения F_m нашего алгоритма следует добавить некоторую новую оценку $h_m(x)$. Таким образом,

$$F_{m+1}(x_i) = F_m(x_i) + h_m(x_i) = y_i.$$

или, что эквивалентно,

$$h_m(x_i) = y_i - F_m(x_i).$$

Таким образом, градиентный бустинг будет соответствовать h_m остатку $y_i - F_m(x_i)$. Как и в других вариантах бустинга, каждый из них F_{m+1} пытается исправить ошибки своего предшественника F_m. Обобщение этой идеи на функции потерь, отличные от квадратичной ошибки, и на проблемы классификации и ранжирования, следует из наблюдения, что невязки $h_m(x_i)$ для данной модели пропорциональны отрицательным градиентам функции потерь среднеквадратичной ошибки (MSE) (относительно $F(x_i)$):

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - F(x_i))^2$$

$$-\frac{\partial L_{MSE}}{\partial F(x_i)} = \frac{2}{n}\left(y_i - F(x_i)\right) = \frac{2}{n}\, h_m(x_i).$$

Итак, градиентный бустинг может быть специализирован на алгоритме градиентного спуска, и его обобщение влечет за собой "подключение" другой потери и ее градиента.

Воспользовавшись технологией Google Colaboratory для использования облачных вычислений с помощью видеокарты удалось выполнить более десяти тысяч итераций и сократить среднеквадратичное отклонение менее 0,2.

График зависимости среднеквадратичного отклонения от первых тысячи итераций представлен на рисунке 2.25.

Рисунок 2.25 – График зависимости среднеквадратичного отклонения от числа итераций.

Благодаря данной технологии удалось построить модель, которая имеет точность предсказаний более 80%.

Применяя функцию сигмоида – это гладкая монотонная возрастающая нелинейная функция, имеющая форму буквы «S», которая часто применяется для «сглаживания» значений некоторой величины, получаем вероятности риска возникновения сердечного приступа у пациента по его признакам, формула сигмоиды и ее график представлены на рисунке 2.26.

$$f(x) = \frac{1}{1 + e^{-x}}$$

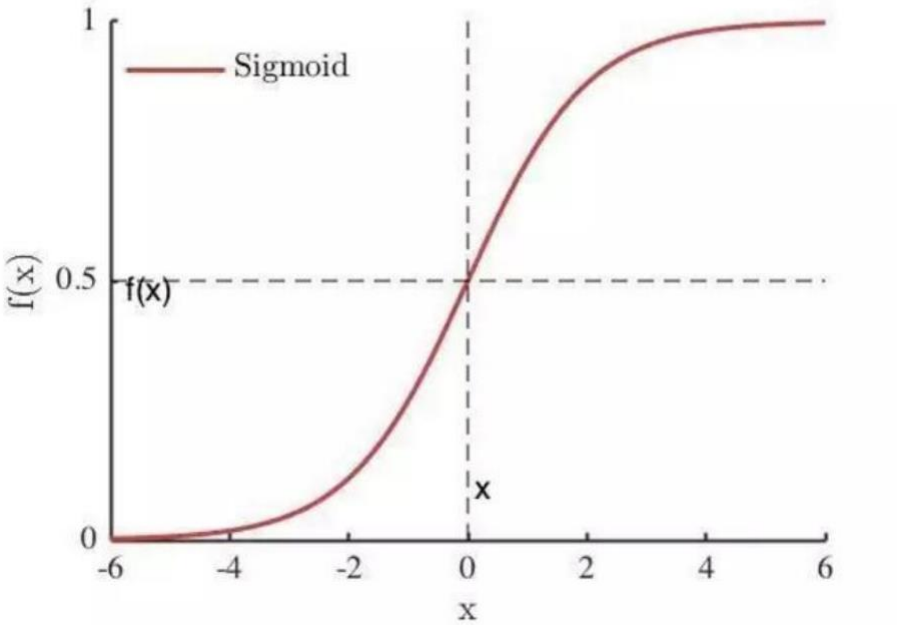

Рисунок 2.26 – Формула и график сигмоиды.

Результаты предсказаний представлены значениями от нуля до одного, на рисунке 2.27.

probabilities_all
0.601591
0.523970
0.545524
0.550629
0.571810
...
0.521147
0.592054
0.604294
0.631161
0.685026

Рисунок 2.27 – Предсказания.

В результате данной курсовой работы получили рабочую модель, которую можно применять для прогнозирования риска сердечного приступа у пациентов.

ЗАКЛЮЧЕНИЕ

Исследование риска сердечного приступа, как одного из важнейших аспектов здоровья человека и анализ основных данных показателей человеческого состояния является важнейшей задачей для всего человечества.

Проверенные и проанализированные данные о связи биологических, демографических и других показателей с риском возникновения инфаркта позволят людям прогнозировать риски возникновения проблем с сердечно-сосудистой системой на ранних этапах развития заболеваний и вовремя защитить себя.

В заключение, важно заметить, что это не идеальное предсказание, а лишь вероятность, для более точных заключений необходимо увеличивать объемы данных не только в размерах, но и в количестве исходных признаков, также улучшать используемую модель и точность вычислений, и тогда возможно человек сможет на 100% прогнозировать риск возникновения сердечного приступа.

Цель данной курсовой работы — построить модель предсказания, проанализировать влияние биологических, демографических и других показателей на риски возникновения инфаркта — достигнута.

В ходе выполнения данной курсовой работы построена модель предсказания, проведен корреляционно-регрессионный анализ влияния признаков на риски возникновения сердечного приступа с использованием языка программирования Python, его библиотек и среды разработки Jupyter Notebook.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1. Инфаркт_миокарда / Свободная энциклопедия Wikipedia [Электронный ресурс]. https://ru.wikipedia.org/wiki/Инфаркт_миокарда

ПРАКТИЧЕСКАЯ ЧАСТЬ

1. Pandas / Документация библиотеки [Электронный ресурс]. https://pandas.pydata.org/docs/
2. Numpy / Документация библиотеки [Электронный ресурс]. https://numpy.org
3. Catboost / Документация библиотеки [Электронный ресурс]. https://catboost.ai
4. Matplotlib / Документация библиотеки [Электронный ресурс]. https://matplotlib.org
5. Scikit-Learn / Документация библиотеки [Электронный ресурс]. https://scikit-learn.org/stable/index.html

ПРИЛОЖЕНИЯ

Приложение А – ссылка.

Приложение А

Ниже представлена ссылка на репозиторий сайта GitHub в котором содержится исходных код данной курсовой работы:

https://github.com/TimmmofeyD/course_2_1/tree/main