© Использование методов машинного обучения в медицине

А.А. Чертилина, e-mail: nastya12.11@yandex.ru
В.В.Мокшин, e-mail: vladimir.mokshin@mail.ru

Казанский национальный исследовательский технический университет им. А. Н. Туполева - Каи

Аннотация. *В данной статье мы рассмотрим что такое машинное обучение, и как оно может помочь в предупреждении инсульта. Мы найдем основную группу риска по собранным данным и построим модель, которая будет помогать строить прогноз заболевания*

Ключевые слова: *инсульт, медицина, данные регрессия, машинное обучение.*

Введение

Ежедневно в сфере медицины генерируется большой объем данных. Это могут быть данные, полученные вследствие проведения теста над новыми медикаментозными средствами. Данные фиксируются и заносятся в таблицу. После проведения анализа, выявляются зависимости между отслеживаемыми параметрами, строятся статистические модели. И на выходе получаем то насколько действенно средство, насколько оно безопасно и какие последствия могут быть.

Так же может проходить сбор биоданных людей, обратившихся в медицинскую службу, совместно с их симптомами и диагнозами. В этом случае методы интеллектуального анализа данных помогут нам найти группу людей наиболее подверженных заболеванию, какие факторы наиболее явно могут сообщить о будущем заражении и так далее.

Так же данные, полученные опытным путем, можно использовать для машинного обучения. Создав тренировочную выборку из сходных данных, мы будем обучать программу предсказывать вероятность того, что человек болен. Что в будущем может помочь врачам ставить более точный диагноз или обнаруживать ранее не известные болезни.

В этой статье будет показана действенность методов машинного обучения в медицине, но относительно простом примере. Таком как вероятность того что человек в ближайшее время перенесет инсульт.

1. Анализ модели

Для исследования берется готовая база данных с сайта Kaggle[1], в которой находится информация о пациентах. База данных содержит следующие столбцы:

1. Пол: "Мужской", "Женский" или "Другой"
2. Возраст: возраст пациента
3. Гипертония: 0, если у пациента нет гипертонии, 1, если у пациента гипертония
4. Сердечные заболевания: 0, если у пациента нет никаких сердечных заболеваний, 1, если у пациента есть заболевание сердца
5. Никогда не был женат: "Нет" или "Да"
6. Тип работы: "ребенок", " работа в правительстве ", "Никогда не работавший", "Частный" или "Самозанятый"
7. Тип местожительства: "Сельский" или "Городской"
8. Средний уровень глюкозы в крови
9. Имт: индекс массы тела
10. Инсульт: 0, если у пациента не было инсульта, 1, если у пациента был инсульт

Основу методов интеллектуального анализа данных составляют все возможные методы классификации, моделирования и прогнозирования. [2] Существует множество методов интеллектуального анализа, но в статье будут рассмотрены статистические методы (корреляционный анализ) и методы моделирования, прогнозирования.

Первым делом посмотрим количество перенесших инсульт и количество тех, кто здоров. Это соотношение можно увидеть на рис.1.

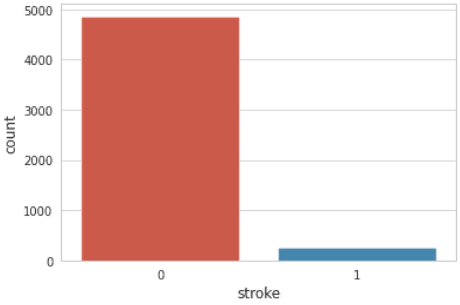

Рис. 1. Соотношение здоровых людей и людей перенесшим инсульт

Как видно из графика, база данных сильно несбалансированная так как людей, перенесших инсульт, слишком мало, в сравнении с теми, у кого его не было. Из-за этого есть достаточно высокая вероятность, что при построении модели получим неточные данные. Поэтому нужно будет выполнить либо избыточную выборку, либо недостаточную.

В эксперименте участвовали как представители мужского пола, так и женского. Разделим выборку по гендерному признаку и посмотрим кто более подвержен риску заболевания. Результат можно увидеть на рисунке 2.

Рис. 2. Соотношение заболевших к здоровым среди мужчин и женщин. Фиолетовым отмечены здоровые люди, розовым- перенесшим инсульт.

Как видно из рисунка, количество участвующих в эксперименте женщин и мужчин разное количество, но стоит отметить, что и та и другая категория подвержена равному риску сердечного инсульта.

Чтобы не перебирать вручную все параметры было бы неплохо увидеть корреляционную связь между разными переменными нашей базы данных. воспользуемся одним из средств интеллектуального анализа данных. На рисунке 3 представлена тепловая матрица, в которой цветом обозначаются значения связи между различными параметрами. Это коэффициенты корреляции. То есть, элемент на пересечении i-й строки и j-ого столбца является коэффициентом корреляции между этой строкой и этим столбцом. Коэффициент определяется по формуле 1:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}}$$

x, y- сравниваемые величины.

r- коэффициент корреляции. Принимает значения от -1 до +1

(xy) ¯, (x) ¯,(y) ¯ -среднее значение признаков.

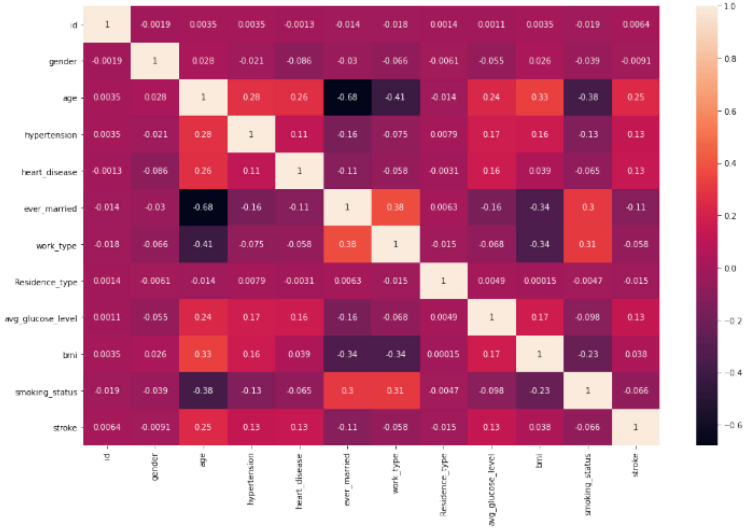

Рис. 3. тепловая карта корреляции

Из матрицы видно, что инсульт сильнее всего связан с возрастом. Немного слабее связан с фактором гипертонии, сердечной болезни и средним уровнем сахара. И совершенно не зависит от гендера, бал ли человек женат, от того какая у него работа.

Все что было выше делалось ради обнаружения группы людей подверженной заболеванию. Так как выло выяснено, что вероятность наступления инсульта зависит от возраста, уровня сахара в крови и индекса массы тела, то построим графики регрессии, чтобы лучше понимать в какой момент человек становится более подвержен инсульту. Эти графики объединим в один рисунок 4.

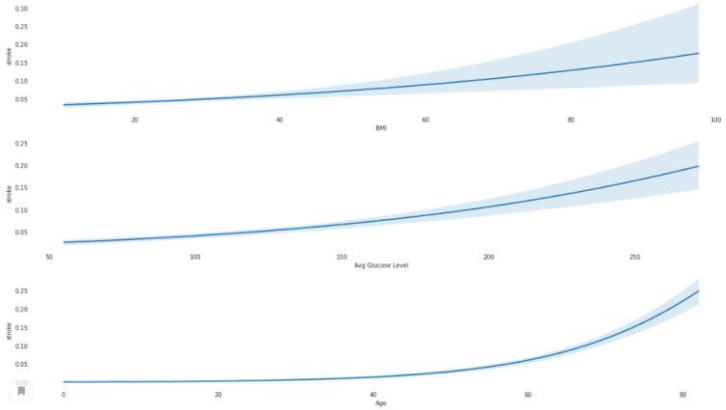

Рис. 4. изменение вероятности заболевания от массы тела, от сахара в крови, от возраста.

И получается, что при просмотре зависимости наступления инсульта от уровня массы тела на графике есть большой разброс по данным. Конечно, чем больше индекс, тем выше вероятность наступления инсульта. Основной наклон начинается с 40 кг/м^2, что соответствует третьей степени ожирения. Перейдем к следующему график. Зависимость уровня сахара в крови и вероятностью наступления инсульта. Здесь разброс куда меньше, но все еще существенный. График более крутой нежели в предыдущем. И можно сделать вывод, что если у пациента уровень сахара в крови выше 300 миллимолей на литер, то повышается вероятность инсульта. Эти данные куда полезнее. Рассмотрим третий график. Разброса почти нет. Это хорошо. Тогда возраст человека является самым главным показателем вероятности наступления инсульта. И риску апоплексического удара подвержены люди старше 40 лет. И чем человек старше, тем выше вероятность наступления инсульта.

Основная группа людей подверженная кровоизлиянию в мозг определена. Это люди чей индекс массы тела выше 40 кг/м^2, уровень сахара выше 300 миллимолей на литер и которым больше 40 лет. Основываясь на этих данных, мы можем обучить искусственный интеллект прогнозировать вероятность апоплексического удара.

2. Машинное обучение

Так как набор данных несбалансирован, то стоит воспользоваться методом SMOTE(метод передискретизации синтетического меньшинства). Таким обзом будут созданы новые

строчки об заболевших. Данный метод не будет дублировать уже созданные записи, он будут генерировать новые наблюдения, но на основании изначальных данных. Этот метод нахоодит несколько ближайших соседей для каждого члена меньшего класса. И тогда один или несколько соседей используются для создания новых наблюдений.

После уравнивания базы данные стоит разделить на обучающую и тестовую выборку. Обучающая выборка предназанчена для построения классификаторов, а тестовая выборка отвечает за оценку качества работы обученного алгоритма.

Теперь можно начать машинное обучение. Так как для обучения программе представлен большой объем данных с ответом был ли у человека инсульт, то обучение относится к типу обучению с учителем. Так как у нас стоит нужда в будущем определить был ли у человека инсульт то эта задача относится к типу классификации. Т.е. есть стремление получения ответа на вопрос либо да, либо нет.

Для обучения модели воспользуемся логической регрессией.[3] Она представляет собой способ определения зависимости между переменными, одни из которых категориально зависимы(инсульт), а другие не зависимы(индекс массы тела, уровень сахара в крови, возраст).[4] И на основе этой регрессии мы испоьзуем алгоритм кривой ROC, который будет менять пороговые значения для нахождения более точных результатов. Точнее, чтобы количество неправильных прогнозов можно было свести к минимуму.

На рисунке 5 нам показано соотношение правельного предсказания к ошибочному.

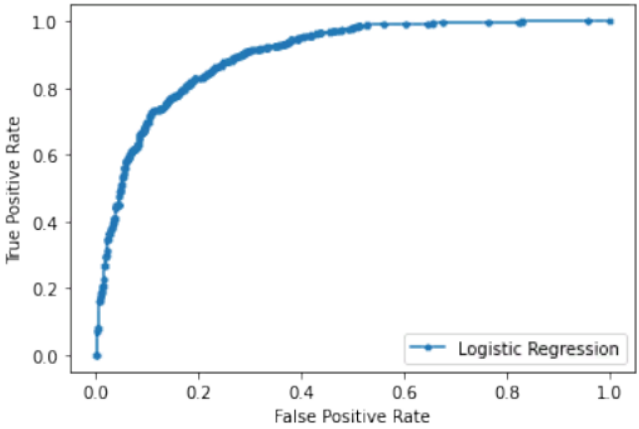

Рис. 5. Кривая ROC.

Модель выдала что правильность ответа или точнее сказать точность предсказания программой равна 81%, что достаточно высоко. Но можно проверить другие методы. [5]

Воспользуемся алгоритмом случайного леса. В этом случае строятся деревья решений на основе параметра для разделения. Таким образом генерируется множество деревьев, которые в совокупности называются лесом. Строим случайный лес и получаем что вероятность получения правильного прогноза равна 92%.

Проверим еще один алгоритм построения модели: классификатор XGBOOST. Метод заключается в создании ансамбля последовательно уточняющих друг друга деревьев решений. После обучения точность предсказания оказалась равной 95%. Не существует алгоритмов способных сделать предсказание со 100% точностью. И на данный момент мы не сможем получит результат выше 95%. В любом случае это очень хороший результат.

Заключение

В данном наборе данных были определены наиболее взаимосвязанные параметры с параметром по названием инсульт. Благодаря чему обнаружены группы людей, находящиеся в зоне риска. Этот результат соответствует основным причинам инсульта. Если расширить собираемые данные, то будет возможность найти дополнительные зависимости наступления инсульта.

После использования машинного обучения и применения разных алгоритмов, получаем что точность правильного прогнозирования равна 95%. И в будущем этот алгоритм может помогать врачам ставить диагноз.

Проведение данной работы возможно не только с данными связанными с инсультом. Можно разработать модель помогающую прогнозировать рак или классифицировать болезнь по биоданным человека. Это может помочь обнаружить болезнь которая до этого еще не была известна и помочь врачам найти подходящее лечение за более короткий срок.

Литература

1. fedesoriano. Stroke Prediction Dataset//26-01-2021 URL: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/metadata (дата обращения 08.12.2021)

2. Дядичев В.В., Ромашка Е.В., Голуб Т.В. Задачи и методы интеллектуального анализа данных // Геополитика и экогеодинамика регионов. 2015. №3. URL: https://cyberleninka.ru/article/n/zadachi-i-metody-intellektualnogo-analiza-dannyh (дата обращения: 20.12.2021).

3. Рекурсивный алгоритм построения регрессионных моделей сложных вероятностных объектов / Мокшин В.В., Сайфудинов И.Р., Кирпичников А.П. Вестник Технологического университета. 2017. Т. 20. № 9. С. 112-116.

4. Якимов И.М. Структурное моделирование бизнес-процессов в системах BPMN EDITOR, ELMA, RUNAWFE /Якимов И.М., Кирпичников А.П., Мокшин В.В., Махмутов М.Т., Пейсахова М.Л., Валиева А.Х., Низамиев Б.А.// Вестник Казанского технологического университета. 2014. Т. 17. № 10. С. 249-256.

5. Метод формирования модели анализа сложной системы /Мокшин В.В., Якимов И.М. Информационные технологии. 2011. № 5. С. 46-51.

6.