Научная статья
Original article
УДК 002.304

# ПРОГНОЗИРОВАНИЕ РИСКА РАЗВИТИЯ СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ С ПОМОЩЬЮ ПОЛНОСВЯЗНОЙ ГЛУБОКОЙ НЕЙРОННОЙ СЕТИ

PREDICTING THE RISK OF DEVELOPING CARDIOVASCULAR DISEASES USING A FULLY CONNECTED DEEP NEURAL NETWORK

**Еникеева З.А.,** кандидат физико-математических наук, доцент кафедры «Программная инженерия» Институт информационных технологий и интеллектуальных систем Казанский (Приволжский) федеральный университет
Россия, г. Казань

**Рафиков Д.С.,** магистр 2 курс, факультет «Окрытая информатика»
Институт вычислительной математики и информационных технологий Казанский (Приволжский) федеральный университет
Россия, г. Казань

**Enikeeva Z.A.,** din.rafikov@gmail.com
**Rafikov D.S.,** din.rafikov@gmail.com

**Аннотация.** задаче прогнозирования риска развития сердечно сосудистых заболеваний. по причине смертности в России многие годы занимают сердечно-сосудистые заболевания (ССЗ). Сотрудниками медицинской сферы были предоставлены обезличенные медицинские данные о 115 тысячах пациентов

больниц республики Татарстан и поставлена задача построения модели, для предсказания риска развития ССЗ.

В статье приведён разведочный анализ данных, конфигурация и обучение полносвязной глубокой нейронной сети, а также построение 95% доверительного интервала для прогнозируемых значений.

**Abstract.** The article is devoted to the task of predicting the risk of cardiovascular disease. cardiovascular disease (CVD) has been the leading cause of mortality in Russia for many years. Employees of the medical field were given anonymized medical data on 115 thousand patients in hospitals of the Republic of Tatarstan and tasked to build a model to predict the risk of CVDs.

An exploratory data analysis, configuration and training of a full-link deep neural network, and construction of a 95% confidence interval for the predicted values are given in the article.

**Ключевые слова:** машинное обучение, прогнозирование значений, нейронная сеть, глубокая нейронная сеть, полносвязная нейронная сеть.

**Keywords**: machine learning, value prediction, neural network, deep neural network, full-link neural network.

## 1. Введение

В современном мире технологии машинного обучения используется во многих сферах жизни. Технологии машинного обучения позволяют анализировать огромные объёмы данных, находить в них определенные тенденции и закономерности, по этой причине всё больше областей, располагающих значительными объёмами данных прибегают к ним. Одной из основных таких направлений является область медицины.

Сотрудниками медицинской сферы были предоставлены обезличенные медицинские данные о пациентах больниц республики Татарстан, которые проходили исследование по программе управление рисками сердечно-сосудистых заболеваний[1]. Задача поставленная коллегами из медицины разработать модель для прогнозирования риска развития сердечно-сосудистых заболеваний по

имеющимся данным. Эталонные оценки рисков присутствуют на части данных, однако нет информации каким образом и на основе каких параметров они вычислялись.

## 2. Разведочный анализ данных

Перед началом работы с данными в первую очередь необходимо внимательно их изучить — провести разведочный анализ данных. Данное определение было предложено математиком Джоном Тьюки и включает в себя[2]:

- максимальное внимание к самим данным;
- поиск и очистка аномалий в данных;
- определение наиболее существенных параметров;
- определение взаимосвязей.

Имеющиеся данные представляют собой таблицу с 28 колонками и 115262 строками, более подробную информацию об данных можно увидеть на рис. 2.1.:

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 115262 entries, 0 to 115261
Data columns (total 28 columns):
 #   Column                                      Non-Null Count   Dtype  
---  ------                                      --------------   -----  
 0   number                                      115262 non-null  int64  
 1   reference_date                              115262 non-null  object 
 2   practitioner_name                           115262 non-null  object 
 3   organisation_name                           115262 non-null  object 
 4   is_city                                     115262 non-null  int64  
 5   treatment_start_date                        115262 non-null  object 
 6   risk_score                                  72439 non-null   float64
 7   mdrd                                        42124 non-null   float64
 8   mdrd_last_measurement_date                  42124 non-null   object 
 9   glucose_fasting                             104119 non-null  float64
 10  glucose_fasting_last_measurement_date       104119 non-null  object 
 11  total_cholesterol                           109147 non-null  float64
 12  total_cholesterol_last_measurement_date     109147 non-null  object 
 13  systolic_blood_pressure                     109718 non-null  float64
 14  diastolic_blood_pressure                    109718 non-null  float64
 15  systolic_blood_pressure_last_measurement_date  109718 non-null  object 
 16  bmi                                         103106 non-null  float64
 17  bmi_last_measurement_date                   103106 non-null  object 
 18  is_primary                                  115262 non-null  object 
 19  smoking_status                              107023 non-null  object 
 20  last_contact_date                           113556 non-null  object 
 21  gender                                      115262 non-null  object 
 22  age                                         115262 non-null  int64  
 23  hypertension                                95878 non-null   object 
 24  annual_checkup_dutch_last_date              26333 non-null   object 
 25  annual_checkup_russian_last_date            97976 non-null   object 
 26  last_interim_checkup_dutch_date             1671 non-null    object 
 27  last_interim_checkup_russian_date           37427 non-null   object 
dtypes: float64(7), int64(3), object(18)
```

Рисунок 2.1. Информация об имеющихся данных загруженных в класс DataFrame библиотеки pandas

Для корректной работы с данными необходимо уточнить типизацию для параметров с типом “object”. Временных параметры были приведены к типу “date”, для числовых “float”. Категориальные типы заменены на новые параметры числового типа. Это касалось параметров “is_primary”, “smoking_status”, “gender” и “hypertension”.

Была проведена проверка на полноту данных по каждому фактору. По причине отсутствия более чем в половине записей “mdrd”, “mdrd_last_measurement_date”, “last_interim_checkup_dutch_date”, “last_interim_checkup_russian_date”, “annual_checkup_dutch_last_date”. Поле “number” уникально для каждой записи и “reference_date” единое для всех, от них не было бы пользы при анализе, поэтому они тоже были удалены.

Следующим шагом является анализ взаимодействия между факторами. Нужно проверить наличие корреляция между параметрами, чтобы оставить только значащие. На рисунке 2.2. можно увидеть матрицу корреляции.

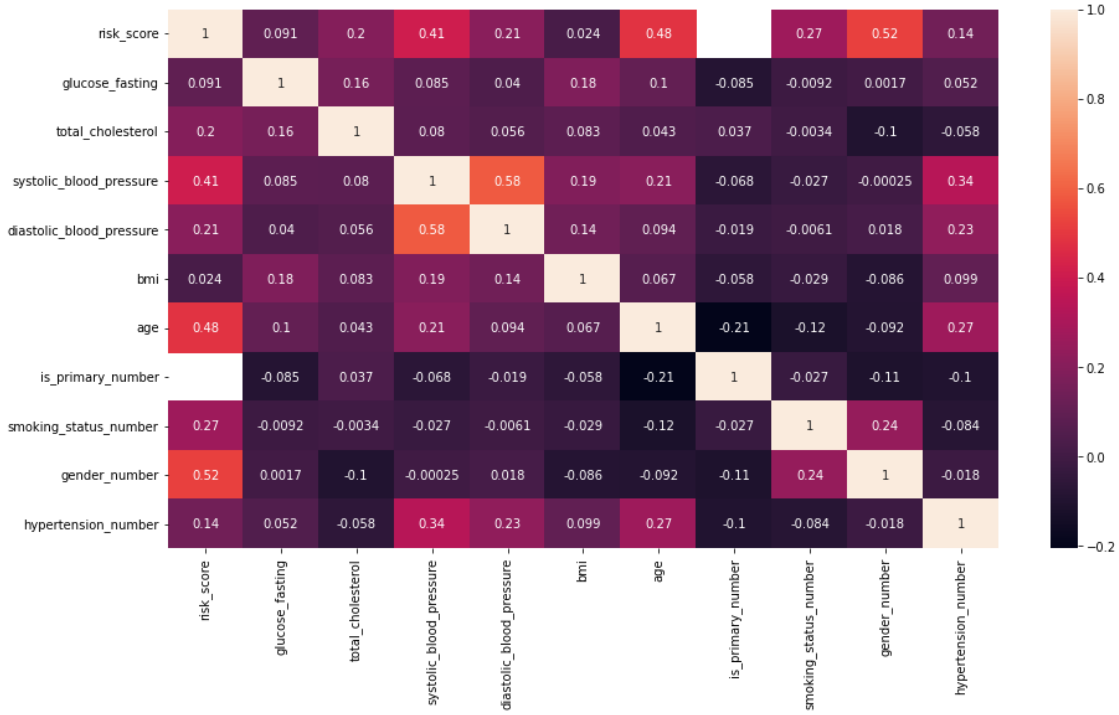


Рис. 2.2. Матрица корреляций параметров

Можно заметить, ожидаемое наличие достаточно значительной корреляции между систолическим и диастолическим уровнем давления. Также корреляцию между риском развития сердечно-сосудистых заболеваний с уровнем давления и возрастом. Однако нет достаточно большого уровня корреляции, чтобы отсеять какой-либо из параметров.

## 3. Конфигурация нейронной сети

В качестве метода классификации с учителем была использована полносвязная глубокая нейронная сеть. Нейронная сеть — математическая модель,

сконструированная по подобию биологических нейронных сетей. Полносвязной называется сеть, в который все нейроны соседних слоев связаны между собой строго в направлении от раннего слоя к последующему. Глубокой нейронной сетью называется модель, где помимо входного и выходного слоя также присутствуют скрытые слои между ними[3].

Хорошей практикой для обучения нейронной сети является предварительная нормализация параметров, поэтому в качестве первого слоя сети был выбран слой нормализации. Т.к. что чем шире область определения параметра, тем сильнее оно будет влиять в начале обучения.

Следующим важным пунктом настройки сети является выбор правильной функции активации для нейронной сети. Для промежуточных слоев лучше всего подходит функция активации ReLU (Rectified Linear units)[4] (3.1). Данная функция активации одновременная простая и эффективная. Основным её преимуществом является решение проблемы исчезающего градиента при обучении сети, чем и обусловлена её популярность.

$$f_{ReLU}(x) = 0 \text{ для } x \leq 0, 1 \text{ для } x > 0 \quad (3.1)$$

Выходной слой сети представляет собой один нейрон, без особой фукницей активации. Так как интересуют конкретные числовые значения, без каких либо трансформаций.

Также значимым пунктом для обучения нейронной сети является выбор функции потерь. Она используется для расчета ошибки между реальными и полученными ответами, т.е. от неё зависит оценка точности модели во время обучения. Подходящими функциями потерь для задачи прогнозирования являются:

- Средняя абсолютная ошибки — вычисляется как среднее значение ошибок предсказаний. Основным преимуществом данной функции является устойчивость к выбросам;
- Среднеквадратическая ошибка — вычисляется как среднее квадратов ошибок предсказаний. Особенностью является более строгое наказание модели, за большие ошибки в прогнозе.

## 4. Обучение нейронной сети

Обучения полносвязной нейронной сети происходило с помощью библиотеки keras. Для прогнозирования, отделяется фактор “risk_score”. Данные были разделены на обучающую и тестовую выборку, последняя составила 20% от общего числа. Информация о выборках можно увидеть на рисунке 4.1.:

`train_dataset.describe().transpose()`

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| is_city | 43897.0 | 0.880789 | 0.324040 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| risk_score | 43897.0 | 3.969975 | 3.243492 | 0.0 | 2.0 | 3.0 | 5.0 | 47.0 |
| glucose_fasting | 43897.0 | 5.288097 | 1.225928 | 3.0 | 4.7 | 5.1 | 5.6 | 25.0 |
| total_cholesterol | 43897.0 | 5.512548 | 0.877794 | 2.7 | 4.9 | 5.4 | 6.1 | 8.8 |
| systolic_blood_pressure | 43897.0 | 137.607057 | 12.160454 | 90.0 | 130.0 | 140.0 | 145.0 | 220.0 |
| diastolic_blood_pressure | 43897.0 | 83.726747 | 7.236847 | 40.0 | 80.0 | 80.0 | 90.0 | 120.0 |
| bmi | 43897.0 | 27.697588 | 4.294835 | 15.0 | 25.0 | 27.0 | 30.0 | 45.0 |
| age | 43897.0 | 63.349887 | 9.809579 | 40.0 | 57.0 | 63.0 | 70.0 | 95.0 |
| is_primary_number | 43897.0 | 1.000000 | 0.000000 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| gender_number | 43897.0 | 0.284849 | 0.451347 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| hypertension_number | 43897.0 | 0.877714 | 0.327620 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| smoked_formely | 43897.0 | 0.010411 | 0.101502 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| smoked_never | 43897.0 | 0.959906 | 0.196182 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| smoking | 43897.0 | 0.029683 | 0.169714 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |

`test_dataset.describe().transpose()`

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| is_city | 10974.0 | 0.877802 | 0.327529 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| risk_score | 10974.0 | 3.950246 | 3.198680 | 0.0 | 2.0 | 3.0 | 5.0 | 29.0 |
| glucose_fasting | 10974.0 | 5.266730 | 1.210181 | 3.0 | 4.7 | 5.1 | 5.6 | 22.5 |
| total_cholesterol | 10974.0 | 5.517387 | 0.876497 | 3.3 | 4.9 | 5.4 | 6.1 | 8.0 |
| systolic_blood_pressure | 10974.0 | 137.540642 | 11.967762 | 90.0 | 130.0 | 140.0 | 145.0 | 180.0 |
| diastolic_blood_pressure | 10974.0 | 83.741844 | 7.189267 | 40.0 | 80.0 | 80.0 | 90.0 | 120.0 |
| bmi | 10974.0 | 27.636140 | 4.250711 | 16.0 | 25.0 | 27.0 | 30.0 | 45.0 |
| age | 10974.0 | 63.170129 | 9.695056 | 40.0 | 57.0 | 63.0 | 69.0 | 95.0 |
| is_primary_number | 10974.0 | 1.000000 | 0.000000 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| gender_number | 10974.0 | 0.286951 | 0.452359 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| hypertension_number | 10974.0 | 0.877255 | 0.328159 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| smoked_formely | 10974.0 | 0.009659 | 0.097810 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| smoked_never | 10974.0 | 0.963732 | 0.186964 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| smoking | 10974.0 | 0.026608 | 0.160943 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |

Рис. 4.1. Статистика по факторам для обучающей и тестовой выборки

Количество скрытых слоев варьируется от 1 до 4. В качестве количества нейронов в слое, выступают числа от 2 до 8, в степени номера слоя по убыванию. Функцией активации промежуточных нейронных сетей выбирается ReLU, а функциями потерь - среднеквадратичное отклонение и абсолютное среднее отклонение.

В качестве оптимизатора модели был выбран Адам, с шагом обучения $step = 10^{i}$, где $i = \underline{1,5}$. Для каждой построенной модели обучение происходило в 15 эпох, по результатам которого была отобрана лучшая модель, по значению функции потерь и точности предсказания и дообучена в 150 эпох.

Лучший результат показала конфигурация с оптимизатором Адам (шагом в 0.001), тремя скрытыми слоями с количеством нейронов 2401, 343 и 7, с функцией активации промежуточных слоев ReLU.

Далее проверка точности предсказания полученной модели на тестовой выборке. На рисунке 4.2. можно увидеть график соотношения предсказанных значений с настоящими.

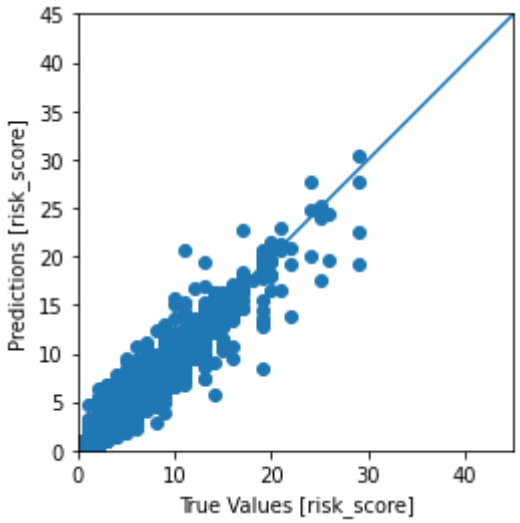


Рисунок 4.2. График соответствия предсказанных точек эталонным

Также проверим распределение отклонений предсказанных значений на рисунке 4.3.. Судя по графику, полученная модель, способна довольно точно предсказывать риск сердечно сосудистых заболеваний по определенному набор факторов, отсутствуют явно выраженные отклонения в одну из сторон.

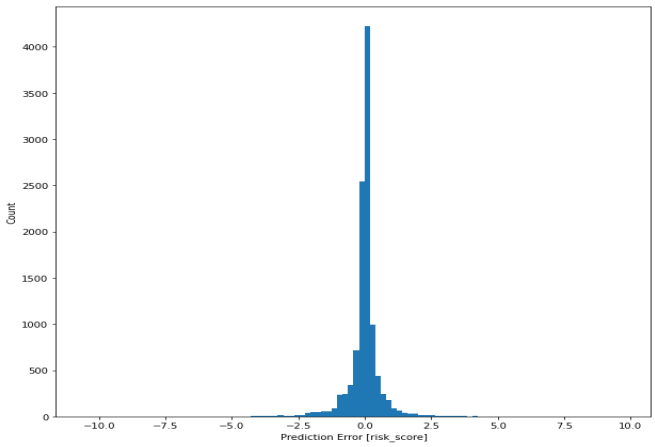


Рис. 4.3. Распределение отклонений предсказанных значений

Для оценки соответствия предсказанных данных к фактическим данным, вычислили коэффициент детерминации и скорректированный коэффициент детерминации[5]. Полученные значения $R^2 = 0.9539$ и $R^2_{adjusted} = 0.9535$. Что свидетельствует о высоком уровне корреляция между предсказанными данными и эталонными.

## 5. Доверительный интервал

Для повышения точности прогнозируемых значений, принято использовать понятие доверительных интервалов. Основная задача данного интервала, покрыть неизвестный фактор, который может быть из-за неточного сбора данных, ошибках в эталонных значения, шума генерируемого нейронной сетью, с определенной надёжностью.

К сожалению, для случая нейронной сети не существует математически обоснованного алгоритма вычисления доверительных интервалов. Поэтому предлагает подход, в котором выбираются границы интерва с помощью квантилей. Для 95% интервала вычисляется левая граница по квантилю 0.025 и правая по 0.975. Полученный интервал [-1.58; 1.18], включает в себе ошибки предсказания для 95% случаев в данной модели. Границы интервала можно увидеть на рисунке 5.1..

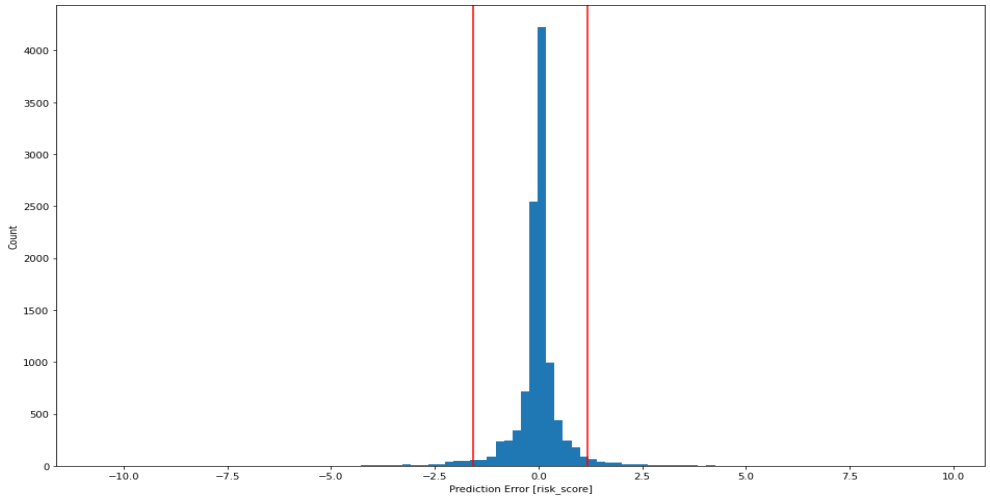


Рис. 5.1. 95% интервал отклонений предсказанных значений

Так как 95% отклонений модели будут попадать в этот интервал, в результате предсказания можно уточнять, что для предсказанного значения x, справедлив доверительный интервал [x - 1.58; x + 1.18].

## Использованные источники:

1. Leif Erhardt, Robert Moller, and Juan García. Comprehensive cardiovascular risk management – what does it mean in practice? / Leif Erhardt, Robert Moller, and Juan García — Vasc Health Risk Manag:Dove Medical Press Limited, 2007. — 603 c.

2. П. Брюс. Разведочный анализ данных / П.Брюс, Э. Брюс. — Практическая статистика для специалистов Data Science. — СПб.: БХВ-Петербург, 2018. — 304 с.
3. Ю. С. Осипов. — Большая российская энциклопедия т.34 / Ю. С. Осипов. — М. : Большая российская энциклопедия, 2017 — 799с.
4. Andrew L. Maas — Rectifier nonlinearities improve neural network acoustic models / Andrew L. Maas, Awni Y. Hannun, Andrew Y. Ng. — Proc. ICML: Computer Science, 2013. — 421с.
5. Бахрушин В. Е. Методы оценивания характеристик нелинейных статистических связей // Системные технологии. — 2011. — № 2(73). — с. 9—14.

**Sources used:**

1. Leif Erhardt, Robert Miller and Juan Garcia. Integrated management of cardiovascular risks – what does this mean in practice? / Leif Erhardt, Robert Moller and Juan Garcia—Vask Health Risk Management: Dove Medical Press Limited, 2007. — 603 p.
2. P. Bruce. Exploratory data analysis / P.Bruce, E. Bruce. — Practical statistics for data science specialists. — St. Petersburg: BHV-Petersburg, 2018. — 304 p.
3. Yu. S. Osipov. — The Great Russian Encyclopedia vol.34 / Y. S. Osipov. — M. : The Great Russian Encyclopedia, 2017 — 799c.
4. Andrew L. Maas — Rectifier nonlinearities improve acoustic models of neural networks / Andrew L. Maas, Auni Ya. Khannun, Andrew Ya. Ng. — Proc. ICML: Informatics, 2013. — 421c.
5. Bakhrushin V. E. Methods for evaluating the characteristics of nonlinear statistical relationships // System technologies. — 2011. — № 2(73). — pp. 9-14.

*© Еникеева З.А., Рафиков Д.С., Научно-образовательный журнал для студентов и преподавателей «StudNet» №6/2022.*

**Для цитирования:** Еникеева З.А., Рафиков Д.С. ПРОГНОЗИРОВАНИЕ РИСКА РАЗВИТИЯ СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ С ПОМОЩЬЮ ПОЛНОСВЯЗНОЙ ГЛУБОКОЙ НЕЙРОННОЙ СЕТИ// Научно-образовательный журнал для студентов и преподавателей «StudNet» №6/2022