Разработка нейронной сети прогнозной модели влияния социально-экономических факторов на состояние психического здоровья населения

Рабинович А.Е., к.э.н., доцент
Московский Политехнический университет, Москва, Россия
Седова Е.А., магистр направления «Системная аналитика больших данных»
Московский Политехнический университет, Москва, Россия

Аннотация. Данная статья посвящена проблеме прогнозирования состояния психического здоровья населения в зависимости от социально-экономического состояния стран с использованием инструментов анализа больших данных. В результате выполнения работы была произведена обработка набора данных по статистике самоубийств в разных странах с 1960 по 2017 годы, а также спроектирована и разработана нейронная сеть для построения прогнозной модели. Разработанная нейронная сеть была протестирована на реальных значениях статистики Российской Федерации за 2018 год.

Ключевые слова: прогнозная модель, нейронные сети, Python, Google Collab, статистика самоубийств, Keras, Numpy

Development of a neural network for a predictive model of the influence of socio-economic factors on the state of mental health of the population

Rabinovich A.E., Candidate of Economic Sciences, Associate Professor
Moscow Polytechnic University, Moscow, Russia
Sedova E.A., master of the direction «System analytics of big data»
Moscow Polytechnic University, Moscow, Russia

Annotation. This article is devoted to the problem of predicting the state of mental health of the population depending on the socio-economic state of countries using big data analysis tools. As a result of the work, the data set on suicide statistics

was processed from 1960 to 2017, and a neural network was designed and developed to build a predictive model. The developed neural network was tested on real values of the statistics of the Russian Federation for 2018.

Keywords: predictive model, neural networks, Python, Google Collab, suicide statistics, Keras, Numpy

Человек, как гражданин, является основой экономического благосостояния своей страны. В процессе выполнения гражданином трудовой деятельности в бюджет страны поступают налоги, которые в дальнейшем расходуются на нужды государства. Каждая страна заинтересована в том, чтобы граждане могли исполнять трудовую деятельность, тем самым увеличивая экономическое благосостояние собственной страны. Возникают ситуации, когда по различным причинам человек не способен вести трудовую деятельность: слишком молодой или преклонный возраст, заболевания или физические травмы, не позволяющие исполнять трудовую деятельность, социально-экономические факторы, безработица и многое другое. В последнее время все чаще возникают случаи, когда физически здоровый человек не может в полной мере осуществлять трудовую деятельность вследствие психических расстройств, которые могут возникать по различными причинам: от генетически передаваемых психических заболеваний до влияния окружающей среды и социально-экономических факторов.

Самым распространенным психическим заболеванием в настоящее время является депрессия [1]. Депрессия – распространенное психическое расстройство и в мировом масштабе одна из основных болезней, приводящих к инвалидности. Около 264 миллионов человек во всем мире страдают от депрессии. Для больных депрессией свойственны подавленное состояние, потеря интереса и способности получать удовольствие, чувство вины, низкая самооценка, нарушения сна и аппетита, утомляемость и плохая концентрация. Депрессия, как долговременная, так и эпизодическая, может существенно

мешать повседневной жизни, работе и учебе. В наиболее тяжелых случаях депрессия может довести до самоубийства.

Оценить и предугадать влияние генетики на состояние психического здоровья населения практически невозможно, однако существуют методы, позволяющие спрогнозировать высшую точку психического расстройства населения – количество самоубийств – в зависимости от социально-экономических факторов страны. Если удастся установить характер взаимосвязи между социально-экономическими факторами и количеством самоубийств, станет возможной разработка мер для их предотвращения.

Целью данной работы является изучение влияния социально-экономических факторов на состояние психического здоровья населения за счет построения и применения модели прогнозирования влияния социально-экономических факторов на состояние психического здоровья населения стран, в частности на количество самоубийств на основе нейронной сети. В рамках данной работы произведена подготовка набора данных для обучения нейронной сети, а также разработана нейронная сеть на основе многослойного персептрона (MLP) для реализации задачи прогнозирования. Построенная нейронная сеть обучена и протестирована на статистике самоубийств в разных странах с 1960 по 2017 годы [2]. При выполнении работы была использована программная среда Python под названием Google Collab.

Перед тем, как приступить к выбору архитектуры нейронной сети и ее построению, необходимо определить к какому классу задач относится поставленная задача. Для этого требуется рассмотреть имеющийся набор данных и параметров, которые будут использоваться в построении прогнозной модели.

Для загрузки набора данных в среду разработки нейронной сети в программной среде Python используется библиотека pandas. На рисунке 1 представлен импортированный набор данных.

```
import pandas as pd
dataset = pd.read_csv('master.csv')
dataset.head()
```

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	796	Boomers

Рис.1 – Импорт набора данных в среду разработки

В рассматриваемом наборе данных представлены следующие ключевые независимые параметры: страна, год, пол, возрастная группа, количество самоубийств, население, значение индекса уровня человеческого развития, годовой ВВП страны и ВВП на душу населения. Остальные параметры являются информационными: количество самоубийств на 100 тыс. населения, страна – год, принадлежность возрастной группы к поколению.

В результате анализа, в представленном выше наборе данных, были выявлены следующие проблемы: наличие специальных символов в заголовках таблицы, неоднородность данных – указание возрастных групп, пола, страны в виде строковых данных, представление колонки «gdp_for_year» в виде строковых данных, наличие неинформативных колонок – «suicides/100k pop», «country-year», «generation», наличие колонки с пустыми значениями (более 50%) – «HDI for year», наличие строковых данных в колонках, которые должны быть использованы для обучения нейронной сети – «country», «sex», «age». Все это делает непригодным применение рассматриваемого набора данных для дальнейшего анализа и обучения нейронной сети. Для решения описанных выше проблем, были произведены следующие манипуляции с набором данных:

1. Был создан датафрейм, в котором были исключены неинформативные колонки и колонки с пустыми значениями;
2. Были исключены специальные символы из заголовка колонок путем замены;
3. Данные в колонке «gdp_for_year» были приведены к числовому типу путем замены символа «,» и преобразования данных при помощи библиотеки pandas, функции pd.to_numeric;

4. При помощи функции библиотеки SkikitLearn – LabelEncoder были закодированы значения колонок «country», «sex», «age».

После предобработки исходных данных, получился новый набор данных, представленный на рисунке 2.

	country	year	sex	age	suicides_no	population	gdp_for_year	gdp_per_capita
0	0	1987	1	0	21	312900	2156624900	796
1	0	1987	1	2	16	308000	2156624900	796
2	0	1987	0	0	14	289700	2156624900	796
3	0	1987	1	5	1	21800	2156624900	796
4	0	1987	1	1	9	274300	2156624900	796
...	...	...	...	...	...	...	...	...
27815	100	2014	0	2	107	3620833	63067077179	2309
27816	100	2014	0	5	9	348465	63067077179	2309
27817	100	2014	1	3	60	2762158	63067077179	2309
27818	100	2014	0	3	44	2631600	63067077179	2309
27819	100	2014	0	4	21	1438935	63067077179	2309

Рис. 2 – Набор данных, подготовленный для обучения нейронной сети

Для того, чтобы производить дальнейшие манипуляции с набором данных, необходимо выделить из датафрейма основной результирующий показатель, который будет прогнозироваться – количество самоубийств (колонку «suicides_no») и привести данные к нормализованному виду.

Нормализация данных производится путем деления каждого значения в колонке на максимальное значение в данной колонке, тем самым все значения в наборе данных принимают значение в диапазоне от 0 до 1. Для нормализации значений набора данных была использована библиотека Numpy. Результат представлен на рисунке 3.

Далее необходимо определить, какой вид нейронной сети будет использоваться для решения поставленной задачи. Существует большое количество видов нейронных сетей для решения задачи прогнозирования, которые могут различаться по степени сложности реализации и настройки, видам задач прогнозирования, функциям и возможностям.

```
[15] for col in X.columns:
       X[col]=X[col]/np.max(X[col])
     X
```

	country	year	sex	age	population	gdp_for_year	gdp_per_capita
0	0.0	0.985615	1.0	0.0	0.007143	0.000119	0.006300
1	0.0	0.985615	1.0	0.4	0.007031	0.000119	0.006300
2	0.0	0.985615	0.0	0.0	0.006613	0.000119	0.006300
3	0.0	0.985615	1.0	1.0	0.000498	0.000119	0.006300
4	0.0	0.985615	1.0	0.2	0.006262	0.000119	0.006300
...	...	...	...	...	...	...	...
27815	1.0	0.999008	0.0	0.4	0.082658	0.003480	0.018274
27816	1.0	0.999008	0.0	1.0	0.007955	0.003480	0.018274
27817	1.0	0.999008	1.0	0.6	0.063055	0.003480	0.018274
27818	1.0	0.999008	0.0	0.6	0.060075	0.003480	0.018274
27819	1.0	0.999008	0.0	0.8	0.032848	0.003480	0.018274

27820 rows × 7 columns

```
y=y/np.max(y)
y
```

```
0        0.000940
1        0.000716
2        0.000627
3        0.000045
4        0.000403
           ...
27815    0.004790
27816    0.000403
27817    0.002686
27818    0.001970
27819    0.000940
Name: suicides_no, Length: 27820, dtype: float64
```

Рис. 3 – Нормализация значений набора данных

В данном случае для построения нейронной сети с целью прогнозирования целочисленного показателя, было решено использовать нейронную сеть в виде многослойного персептрона (MLP). Многослойные персептроны являются классическим видом нейронной сети. Они состоят из одного или нескольких слоев нейронов. Данные поступают на входной слой, может быть один или несколько скрытых слоев, обеспечивающих уровни абстракции, и на выходном уровне делаются прогнозы, также называемые видимым уровнем [3]. MLP подходят для задач прогнозирования регрессии, когда реальное значение прогнозируется с учетом набора входных данных. Данные часто предоставляются в табличном формате, например, в CSV-файле или электронной таблице.

Перед началом конфигурации нейронной сети необходимо подготовить обучающую и тестовую выборки данных [4]. Для подготовки обучающей и тестовой выборок, изначальный набор данных был разделен случайным образом. Доля тестовых данных составила 20% к общему числу наблюдений (рисунок 4).

```
test_size = 0.2
seed = 7
X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size=test_size, random_state=seed)
```

Рис. 4 – Подготовка тестовой и обучающей выборки для обучения нейронной сети

Далее с использованием библиотеки Keras была произведена конфигурация нейронной сети. Число входных параметров – 7, один входной слой, 3 скрытых слоя и один выходной. В качестве функции активации для скрытых слоев применяется функция «relu» (rectified linear unit), для выходного слоя применяется функция активации «linear» (см. рисунок 5).

```
m=7
model = keras.Sequential(
    [
        keras.Input(shape=(m,)),
        layers.Dense(16, activation='relu'),
        layers.Dense(8, activation='relu'),
        layers.Dense(4, activation='relu'),
        layers.Dense(1, activation='linear')
    ])
model.summary()
```

```
Model: "sequential_22"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_88 (Dense)             (None, 16)                128
_________________________________________________________________
dense_89 (Dense)             (None, 8)                 136
_________________________________________________________________
dense_90 (Dense)             (None, 4)                 36
_________________________________________________________________
dense_91 (Dense)             (None, 1)                 5
=================================================================
Total params: 305
Trainable params: 305
Non-trainable params: 0
```

Рис. 5 – Конфигурация нейронной сети

Далее нейронная сеть была скомпилирована с функцией потерь и используемыми метриками «mse» (mean squared error) и оптимизатором Nadam. Для оптимизации процесса обучения была добавлена функция «callbacks», которая прерывает процесс обучения в случае, если в течение пяти эпох обучения не будет замечено улучшений. Обучение производилось на протяжении 30 эпох.

Для оценки качества обучения модели была импортирована функция plot_history. Одним из стандартных обратных вызовов, который регистрируется при обучении всех моделей глубокого обучения, является история обратного вызова, которая записывает показатели обучения для каждой эпохи. Она включает в себя потери и точность (для задач классификации), а также потери и точность для набора данных проверки, если он установлен [5]. В результате выполнения функции были построены графики потерь и среднеквадратической ошибки прогнозов, отображенные на рисунке 6.

```
loss, accuracy = model.evaluate(X_train, Y_train, verbose=False)
print('Training Accuracy: {:.4f}'.format(accuracy))
loss, accuracy = model.evaluate(X_test, Y_test, verbose=False)
print('Testing Accuracy: {:.4f}'.format(accuracy))
plot_history(h.history)
```

```
Training Accuracy: 0.0003
Testing Accuracy: 0.0003
```

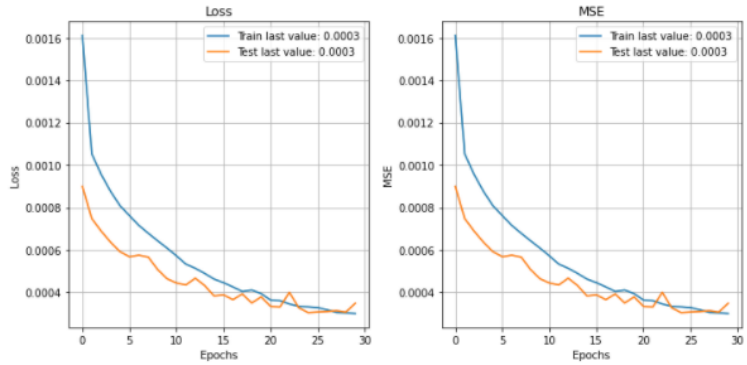

Рис. 6 – Графики с оценкой обучения нейронной сети

Из приведенных графиков видно, что на первых эпохах различия между тренировочными и тестовыми данными были значительными, однако в процессе обучения нейронная сеть научилась прогнозировать значения с небольшой погрешностью. В данном случае можно сказать, что количество эпох для обучения было подобрано верно, и нейронная сеть не успела переобучиться.

Далее необходимо проверить разработанную нейронную сеть и протестировать выдаваемые ей значения прогнозов. Для того, чтобы подать новые значения в нейронную сеть и получить прогноз, необходимо подготовить датафрейм, в который будет загружен новый набор данных. Он должен состоять из тех же колонок, что и набор данных, используемый для обучения нейронной сети. Также, строковые данные должны подаваться в закодированном виде, и все данные, подаваемые в нейронную сеть, должны быть нормализованы.

В результате был подготовлен датафрейм, в который были загружены данные, по которым необходимо произвести прогнозирование показателей. В качестве проверки были выбраны экономические показатели Российской Федерации в 2018 году, и была поставлена задача прогнозирования количества самоубийств для мужчин в возрастной группе от 35 до 54 лет. Данные были предварительно нормализованы для подачи на прогнозирование в нейронную сеть. Результат подготовки данных для прогнозирования представлен на рисунке 7.

```
[67] for col in df.columns:
       df[col]=df[col]/np.max(X_copy[col])
     df
```

	country	year	sex	age	population	gdp_for_year	gdp_per_capita
0	0.84	1.000992	1.0	0.6	0.065808	0.000833	119507.407876

Рис. 7 – Датафрейм для прогнозирования значений с помощью нейронной сети

В результате выполнения программы было получено нормализованное значение количества самоубийств, которое было умножено на максимальное значение столбца в первоначальном наборе данных, для получения денормализованного значения. Нейронная сеть спрогнозировала примерное значение количества самоубийств для выбранных значений страны и возрастной группы. Результат представлен на рисунке 8.

```
[74] ynew = model.predict(df)

     ynew1 = ynew[0][0] * maxy
     ynew1

     8328.098324
```

Рис. 8 – Прогнозное значение нейронной сети

В результате выполнения прогнозирования, было получено примерное значение в 8328 самоубийств на 2018 год для Российской Федерации для мужчин в возрастной группе 35-54 года. По подтвержденным данным общее количество самоубийств в 2018 году составило 18159 человек [6]. Это позволяет сделать вывод, что исходя из прогнозируемого значения количества самоубийств,

примерно 45% от общего количества самоубийств в 2018 году совершается мужчинами в возрасте от 35 до 54 лет.

В результате выполнения исследования была спроектирована и разработана нейронная сеть, прогнозирующая количество самоубийств по введенным параметрам страны (годовой ВВП, ВВП на душу населения), году, полу и возрастной группе. Была произведена оценка потерь при обучении нейронной сети вводом функции потерь и отображением графика обучения. Также предложенная нейронная сеть позволила получить прогнозные значения, согласующиеся с реальными статистическими данных по Российской Федерации за 2018 год.

Разработанная нейронная сеть, а также описанные стадии подготовки данных для обучения, предлагают алгоритм построения простой прогнозной модели, которая может быть использована для реализации более сложной модели прогнозирования, которая впоследствии может быть использована для разработки профилактических социально-экономических мероприятий для превентивного предотвращения увеличения статистики самоубийств.

Библиографический список

1. Психические расстройства // Всемирная организация здравоохранения URL: https://www.who.int/ru/news-room/fact-sheets/detail/m. (дата обращения: 07.02.2021);

2. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018 // http://dx.doi.org/10.1016/S0140-6736(18)32203-7 (дата обращения: 08.02.2021);

3. Гафаров Ф.М Г12 Искусственные нейронные сети и приложения: учеб. пособие / Ф.М. Гафаров, А.Ф. Галимянов. – Казань: Изд-во Казан. ун-та, 2018. – 121 с.;

4. Решение задачи регрессии полносвязной нейронной сетью // Жизнь, бизнес и IT URL: http://www.bizkit.ru/2019/11/05/14921/ (дата обращения: 15.02.2021);

5. Показать историю обучения модели глубокого обучения в Keras // Машинное обучение, нейронные сети, искусственный интеллект URL: https://www.machinelearningmastery.ru/display-deep-learning-model-training-history-in-keras/ (дата обращения: 19.02.2021);

6. Самоубийства в России за год статистика, количество по регионам // STATDATA URL: http://www.statdata.ru/samoubijstva-v-rossii (дата обращения: 21.02.2021).

References

1. Mental disorders // World Health Organization URL: https://www.who.int/ru/news-room/fact-sheets/detail/m (date accessed: 07.02.2021);

2. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. The

Lancet. 2018 // http://dx.doi.org/10.1016/S0140-6736(18)32203-7 (date accessed: 08.02.2021);

3. Gafarov FM G12 Artificial neural networks and applications: textbook. manual / F.M. Gafarov, A.F. Galimyanov. - Kazan: Kazan Publishing House. un-ta, 2018. –121 p.;

4. Solution of the regression problem by a fully connected neural network // Life, business and IT URL: http://www.bizkit.ru/2019/11/05/14921/ (date of access: 15.02.2021);

5. Show the history of training a deep learning model in Keras // Machine learning, neural networks, artificial intelligence URL: https://www.machinelearningmastery.ru/display-deep-learning-model-training-history-in-keras/ (date access: 19.02.2021);

6. Suicides in Russia for the year statistics, number by region // STATDATA URL: http://www.statdata.ru/samoubijstva-v-rossii (date of access: 21.02.2021).