МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

УО «БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра математических методов в экономике

Специальность «Экономическая кибернетика»

**ПРЕЗЕНТАЦИЯ**

**ДИПЛОМНОЙ РАБОТЫ**

на тему: **Разработка моделей нейронных сетей и их использование при принятии решений о выдаче кредита (на примере ОАО «Белинвестбанк»)**

|  |  |  |
| --- | --- | --- |
| Студент  ФЦЭ, 4-й курс, ДКК-1 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Ф.А. Кобак |
| Руководитель  доктор экон. наук,  профессор | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Э.М. Аксень |

МИНСК 2022

**РЕФЕРАТ**

Дипломная работа: 112 c., 16 табл., 33 рис., 7 лист., 25 ист., 11 прил.

КЛАССИФИКАЦИЯ, КРЕДИТНЫЙ СКОРИНГ, ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ, АЛГОРИМ ОБРАТНОГО РАСПРОСТРАНЕНИЯ ОШИБКИ, ВАЛИДАЦИЯ МОДЕЛИ, ОТБОР ПОКАЗАТЕЛЕЙ, AUC, KS, ТЕСТ КОЛМОГОРОВА-СМИРНОВА

**Объект исследования** – правило классификации кредитополучателей в ОАО «Белинвестбанк».

**Предмет исследования –** модели искусственных нейронных сетей в приложении к задаче классификации.

**Цель работы:** построение нелинейного классификатора для целей кредитного скоринга.

**Методы исследования:** компьютерный анализ данных, машинное обучение.

**Исследования и разработки:** при отборе показателей и преобразованиях данных для модели задействован особый метод, использующий ROC анализ, подготовлен ряд инструментов для автоматизации процесса подготовки данных.

**Элементы научной новизны:** новый метод оценки индивидуальной классифицирующей способности каждого показателя.

**Область возможного практического применения:** задачи требующие проведения классификации клиентов по собранным данным.

**Технико-экономическая и социальная значимость:** построение моделей, подобных рассмотренным в работе, позволит решить задачи в которых производительности классических методов классификации недостаточно.

Автор работы подтверждает, что приведенный в ней расчетно-аналитический материал правильно и объективно отражает состояние исследуемого процесса, а все заимствованные из литературных и других источников теоритические, методологические и методические положения и концепции сопровождаются ссылками на их авторов.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 5](#_Toc104015768)

[1 Торические основы моделей нейронной сети 7](#_Toc104015769)

[1.1 Постановка задачи 7](#_Toc104015770)

[1.2 Модель логистической регрессии – линейный классификатор 9](#_Toc104015771)

[1.3 Модель искусственной нейронной сети в классификации 16](#_Toc104015772)

[1.4 Целевые функции и алгоритм обратного распространения ошибки 22](#_Toc104015773)

[2 Экономический анализ кредитоспособности заёмщиков ОАО «Белинветбанк» 28](#_Toc104015774)

[2.1 Начальный анализ наблюденных данных 28](#_Toc104015775)

[2.2 Начальный анализ наблюденных данных 30](#_Toc104015776)

[2.3 Математические методы отбора показателей 37](#_Toc104015777)

[3 Построение и валидация модели 50](#_Toc104015778)

[3.1 Программное описание модели и алгоритма обучения 50](#_Toc104015779)

[3.1.1 Основные элементы модели в pytorch 50](#_Toc104015780)

[3.1.2 Реализация алгоритма обучения 52](#_Toc104015781)

[3.2 Обучение и валидация модели 56](#_Toc104015782)

[3.1.1 Подбор параметров обучения 56](#_Toc104015783)

[3.2.2 Более тонкая подгонка и валидация финальной модели 60](#_Toc104015784)

[ЗАКЛЮЧЕНИЕ 67](#_Toc104015785)

[ПРИЛОЖЕНИЕ А Визуализация сигмоиды двух переменных 72](#_Toc104015786)

[ПРИЛОЖЕНИЕ Б Архитектуры с сигмоидой на выходном слое 73](#_Toc104015787)

[ПРИЛОЖЕНИЕ В Бухгалтерский баланс банка за 2016-2020 гг. 74](#_Toc104015788)

[ПРИЛОЖЕНИЕ В Обобщенные программные функции процессинга данных 76](#_Toc104015789)

[ПРИЛОЖЕНИЕ Д Данные на разных этапах обработки 79](#_Toc104015790)

[ПРИЛОЖЕНИЕ Е Графическая интерпретация TP, FP 92](#_Toc104015791)

[ПРИЛОЖЕНИЕ Ж ROC анализ для номинативной переменной 94](#_Toc104015792)

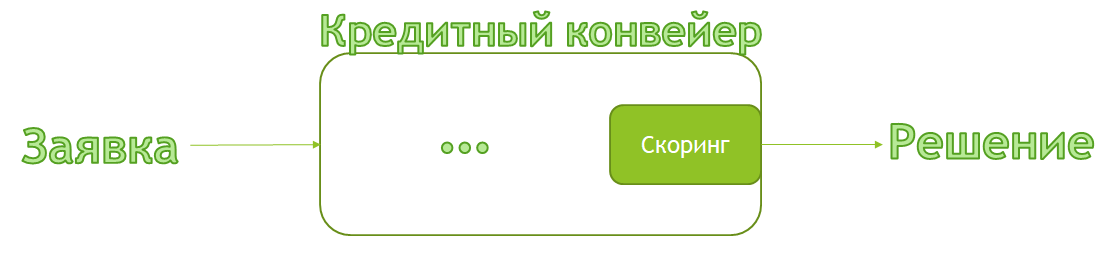
[ПРИЛОЖЕНИЕ И ROC анализ данных для модели 96](#_Toc104015793)

[ПРИЛОЖЕНИЕ К Алгоритм оценки параметров 105](#_Toc104015794)

[ПРИЛОЖЕНИЕ Л Исследование обучения по эпохам 110](#_Toc104015795)

[ПРИЛОЖЕНИЕ М Характеристики модели при обучении 111](#_Toc104015796)

* **Постановка задачи и идентификационная форма модели**



**Рисунок 1 – Место скоринга в принятии решения о выдаче кредита.**

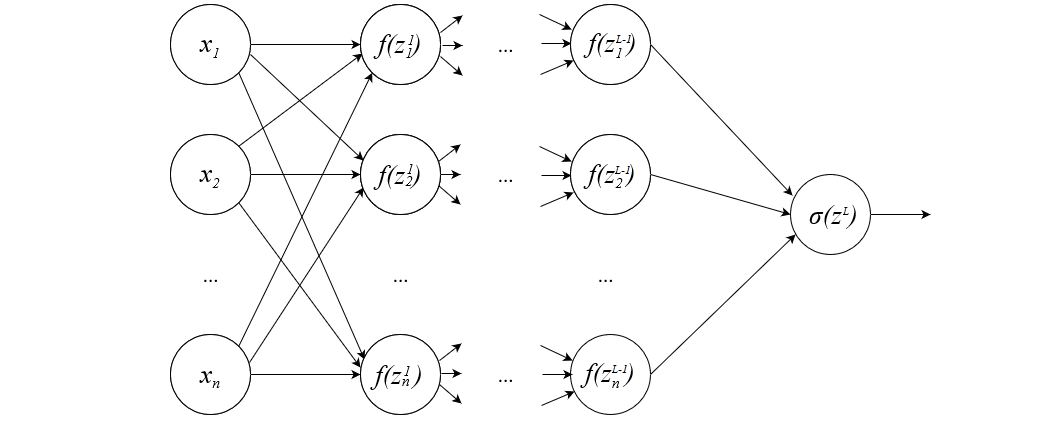
Примечание ­­­– Источник: собственная разработка.

**Таблица 1 – Идея преодоления линейности классификатора**

|  |  |
| --- | --- |
| Логистическая регрессия – линейный классификатор | Нейронная сеть – нелинейный классификатор |
|  |  |
|  |  |

Примечания:

1. Источник: собственная разработка,
2. – функция подставляемая в логит,
3. – вектор характеристик некоторого заёмщика,
4. – область точек, для которой используется соответствующее уравнение ,
5. – коэффициенты модели ,;

****

**Рисунок 2 – Идентификационная форма модели в виде графа**

Примечания:

1. Источник: собственная разработка,
2. – сумматорная функция *i-го* нейрона *l-го* слоя ;,
3. – активация *j-го* нейрона *(l-1)-го* слоя ,
4. , – коэффициенты *l*-*го* слоя модели,
5. – число нейронов в *l*-м слое,
6. – число слоев в модели,
7. – ReLU преобразование,
8. – логит преобразование;

* **Набор данных и его подготовка**

**Таблица 2 – Начальный формат данных.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Название показателя | Тип данных | Область допустимых значений | Число допустимых значений | Число пропусков |
| 1 | 2 | 3 | 4 | 5 |
| Автомобиль год выпуска1 | Дата | [18.01.1986;22.12.2019] | - | 202744 |
| Автомобиль год выпуска2 | Дата | [22.01.1986;10.11.2019] | - | 245403 |
| Автомобиль год выпуска3 | Дата | [11.03.1987;30.05.2019] | - | 246960 |
| Воинская служба | Номинативная | военнослужащий, другое, не отслужил, невоеннообязанный, отсрочка, призывник, уволен в запас | 7 | 108914 |
| Количество детей | Действительное число | [1.0;10.0] | - | 161319 |

**Продолжение таблицы 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| Количество иждивенцев | Действительное число | [1.0;10.0] | - | 237508 |
| Недвижимость | Номинативная | есть, нет | 2 | 40 |
| Образование | Номинативная | высшее, незаконченное высшее, неполное среднее, среднее, среднее специальное | 5 | 70933 |
| Отношение к банку | Номинативная | акционер, другое, не имеет отношения, работник | 4 | 1060 |
| Работа занимаемая должность | Номинативная | безработный (временно не работающий), государственный служащий, заместитель руководителя, индивидуальный предприниматель, пенсионер, рабочий, руководитель, специалист, студент | 9 | 0 |
| Работа последнее место стаж лет | Действительное число | [-998.0;50.0] | - | 1035 |
| Работа уровень дохода BYR | Действительное число | [-1500.0;92207.0] | - | 923 |
| Семейное положение | Номинативная | вдовец/вдова, женат/замужем, повторный брак, разведен/разведена, холост/не замужем | 5 | 63178 |
| Собственная квартира | Номинативная | есть, нет | 2 | 40 |
| Собственный дом | Номинативная | есть, нет | 2 | 40 |
| Уголовная ответственность | Номинативная | есть, нет | 2 | 40 |
| Адрес проживания - Населенный пункт | Номинативная | Название населенного пункта | 6122 | 0 |
| Адрес проживания - Тип населенного пункта | Номинативная | Агрогородок, Город, Городcкой поселок, Деревня, Курортный поселок, Поселок сельского типа, Рабочий поселок, Село, Сельский населенный пункт, Хутор | 10 | 80333 |
| Вид деятельности по ОКЭД | Номинативная | Вид деятельности по ОКЭД | 44 | 15253 |

**Продолжение таблицы 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| Гражданин РБ | Номинативная | Без гражданства, Другое, РБ | 3 | 68942 |
| Дата регистрации договора | Дата | [04.01.2016;21.12.2020] | - | 0 |
| Дата рождения | Дата | [03.01.1878;30.09.2002] | - | 0 |
| Был ли хоть один договор прекращен досрочно | Номинативная | есть, нет | 2 | 137919 |
| Количество фактов просрочки по основному долгу | Действительное число | [0.0;284.0] | - | 165097 |
| Максимальное количество дней просрочки | Действительное число | [0.0;4471.0] | - | 172113 |
| Максимальный срок, на который заключался договор, в годах | Действительное число | [0.0;122.08] | - | 139103 |
| Наличие кредитной истории | Номинативная | есть, нет | 2 | 134471 |
| Общее количество запросов в КБ | Действительное число | [0.0;222.0] | - | 135276 |
| Отношение факт срока к плановому при прекращении КД | Действительное число | [0.0;10000.0] | - | 145229 |
| Причина прекращения договора | Номинативная | Прекращение договора исполнением, Прекращение договора по иным основаниям предусмотренным законодательством Республики Беларусь, Прекращение договора по решению суда, Прощение долга | 4 | 151260 |
| Сумма кредитных лимитов | Действительное число | [0.0;678562.6] | - | 149948 |
| Дата планируемого закрытия | Дата | [20.05.2016;20.07.2058] | - | 10566 |

**Окончание таблицы 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| Дата фактического закрытия | Дата | [15.01.2016;08.07.2021] | - | 107885 |
| Кредитный продукт | Номинативная | Название кредитного продукта | 262 | 748 |
| Овердрафт | Номинативная | есть, нет | 2 | 0 |
| Сумма договора | Действительное число | [-549.81;600000.0] | - | 0 |
| Количество потребляемых банковских продуктов | Номинативная | 1, 2, более 2-х | 3 | 118779 |
| Место работы | Номинативная | Наименование предприятия | 123712 | 1445 |
| Пол | Номинативная | Ж, М | 2 | 0 |
| Социальная группа | Номинативная | безработный, индивидуальный предприниматель, пенсионер, работающий по найму, служащий, учащийся | 6 | 89495 |
| Дефолт | Целое число | [0;1693] | - | 0 |
| Код подразделения | Номинативная | 739-100, 739-200, 739-200-202, 739-200-203, 739-200-228, 739-300, 739-400, 739-600, 739-800, 739-800-831, 739-900, 739-900-500, 739-900-527, 739-900-535, 739-900-536, 739-900-537, 739-900-538, 739-900-905, 739-900-906, 739-900-907, 739-900-932, 739-900-933 | 22 | 0 |

Примечание ­– Источник: собственная разработка.

**Таблица 3 – Описание показателей в которых обнаружены выбросы.**

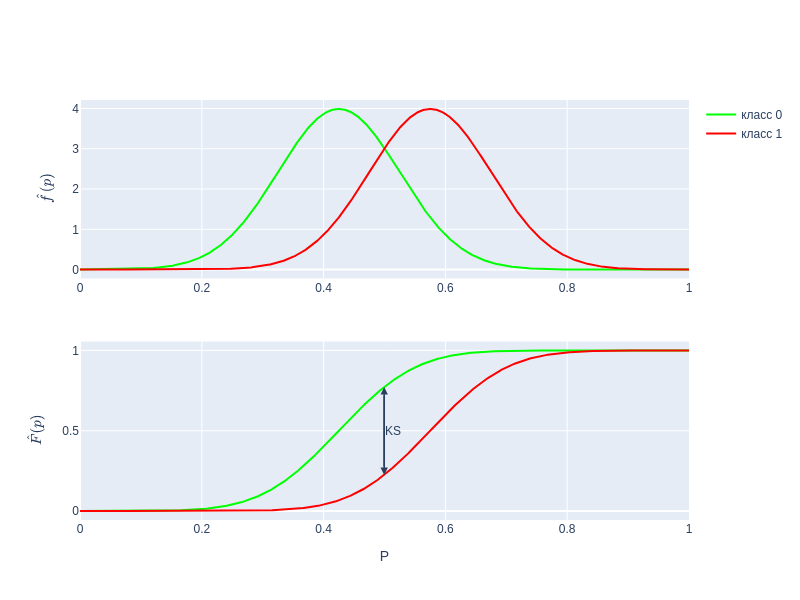
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ста-тистика | Количество фактов просрочки по основному долгу | Макси-мальное количество дней просрочки | Общее коли-чество запросов в КБ | Сумма креди-тных лимитов | Сумма договора | Ежедне-вный платеж |
| Среднее | 5,11 | 18,05 | 8,35 | 7719,67 | 5739,18 | 2,69 |
| СКО | 10,60 | 132,79 | 7,48 | 13222,99 | 13752,80 | 2,74 |
| Мин. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 25% | 0,00 | 0,00 | 3,00 | 1187,16 | 600,00 | 1,14 |
| 50% | 1,00 | 1,00 | 7,00 | 3800,00 | 1500,00 | 1,86 |
| 75% | 5,00 | 8,00 | 11,00 | 8900,00 | 5000,00 | 3,35 |
| Макс. | 284,00 | 4471,00 | 222,00 | 678562,60 | 600000,00 | 196,24 |

Примечание ­– Источник: собственная разработка.

**Таблица 4 – Описание показателей после обработки выбросов.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ста-тистика | Количество фактов просрочки по основному долгу | Макси-мальное количество дней просрочки | Общее коли-чество запросов в КБ | Сумма креди-тных лимитов | Сумма договора | Ежедне-вный платеж |
| Среднее | 2,04 | 3,04 | 7,42 | 4684,76 | 2667,00 | 2,25 |
| СКО | 2,89 | 4,18 | 5,28 | 4643,93 | 2931,37 | 1,48 |
| Мин. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 25% | 0,00 | 0,00 | 3,00 | 1000,00 | 570,80 | 1,12 |
| 50% | 1,00 | 1,00 | 6,00 | 3200,00 | 1118,88 | 1,76 |
| 75% | 3,00 | 8,00 | 11,00 | 7000,00 | 4000,00 | 3,01 |
| Макс. | 12,00 | 20,00 | 23,00 | 20464,32 | 11600,00 | 6,67 |

Примечание ­– Источник: собственная разработка.



**Рисунок 3 – Идея оценки классифицирующей способности показателя**

1. Источник: собственная разработка,
2. – эмпирическая функция плотности вероятности,
3. – эмпирическая функция распределения,
4. – статистика Колмогорова-Смирнова;

**Таблица 5 – Система показателей используемая при построении модели**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Название показателя | Тип данных | Область допустимых значений | Число допустимых значений | Число пропусков |
| 1 | 2 | 3 | 4 | 5 |
| Воинская служба | Номинативная | военнослужащий, невоеннообязанный, уволен в запас, не отслужил, отсрочка, призывник, другое, нет данных | 8 | 0 |
| Недвижимость | Номинативная | есть, нет | 2 | 0 |
| Образование | Номинативная | среднее специальное, высшее, среднее, незаконченное высшее, неполное среднее, нет данных | 6 | 0 |

**Продолжение таблицы 5**

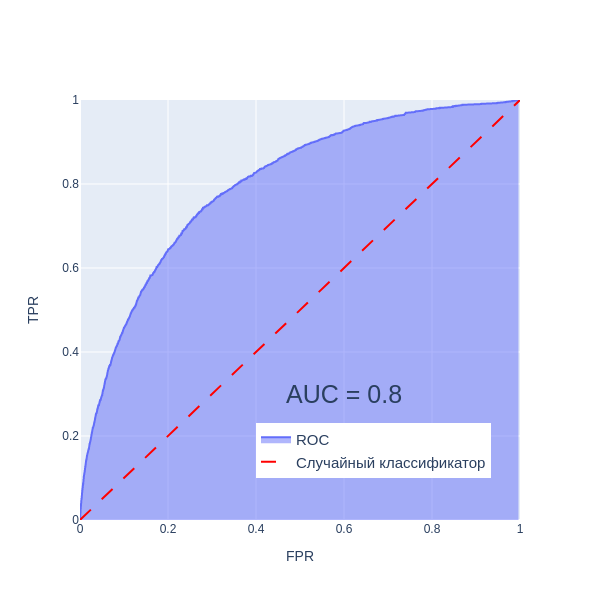
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| Работа занимаемая должность | Номинативная | специалист, рабочий, руководитель, заместитель руководителя, государственный служащий, пенсионер, студент, индивидуальный предприниматель | 8 | 0 |
| Семейное положение | Номинативная | женат/замужем, холост/не замужем, разведен/разведена, вдовец/вдова, повторный брак, нет данных | 6 | 0 |
| Собственная квартира | Номинативная | есть, нет | 2 | 0 |
| Собственный дом | Номинативная | нет, есть | 2 | 0 |
| Уголовная ответственность | Номинативная | нет, есть | 2 | 0 |
| Адрес проживания - Тип населенного пункта | Номинативная | Город, Деревня, Городcкой поселок, Агрогородок, Рабочий поселок, Поселок сельского типа, Сельский населенный пункт, нет данных, Курортный поселок, Село, Хутор | 11 | 0 |
| Гражданин РБ | Номинативная | РБ, Другое, Без гражданства, нет данных | 4 | 0 |
| Дата рождения | Действительное число | [-41611.0;3949.0] | - | 0 |
| Был ли хоть один договор прекращен досрочно | Номинативная | нет данных, есть, нет | 3 | 0 |
| Наличие кредитной истории | Номинативная | нет данных, есть, нет | 3 | 0 |
| Сумма договора | Действительное число | [0.0;11600.0] | - | 0 |
| Количество потребляемых банковских продуктов | Номинативная | более 2-х, 1, 2, нет данных | 4 | 0 |
| Пол | Номинативная | М, Ж | 2 | 0 |

**Окончание таблицы 5**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| Социальная группа | Номинативная | служащий, работающий по найму, нет данных, учащийся, пенсионер, безработный, индивидуальный предприниматель | 7 | 0 |
| Код подразделения | Номинативная | 739-600, 739-900, 739-900-500, 739-400, 739-900-536, 739-300, 739-900-932, 739-200-202, 739-800, 739-200, 739-800-831, 739-900-538, 739-900-905, 739-900-906, 739-900-535, 739-900-537, 739-200-228, 739-900-527, 739-200-203, 739-900-933, 739-900-907, 739-100 | 22 | 0 |
| Y | Целое число | [0;1] | - | 0 |
| Количество действующих кредитных договоров | Действительное число | [0.0;26.0] | - | 0 |
| Количество фактов просрочки по основному долгу | Действительное число | [0.0;12.0] | - | 0 |
| Максимальное количество дней просрочки | Действительное число | [0.0;20.0] | - | 0 |
| Сумма кредитных лимитов | Действительное число | [0.0;20464.32] | - | 0 |
| Автомобиль год выпуска1 | Номинативная | <=3173.0, нет данных, >3173.0 | 3 | 0 |
| Количество запросов в КБ за последние 30 дней | Номинативная | нет данных, <=2.0, >2.0 | 3 | 0 |
| Максимальный срок, на который заключался договор, в годах | Номинативная | нет данных, <=7.94, >7.94 | 3 | 0 |
| Общее количество запросов в КБ | Номинативная | нет данных, <=12.0, >12.0 | 3 | 0 |
| Срок кредита в днях | Номинативная | <=934.0, >934.0, нет данных | 3 | 0 |

Примечание ­– Источник: собственная разработка.

* **Описание модели и выбор точки отсечения**



**Рисунок 4 – ROC кривая предложенного классификатора**

1. Источник: собственная разработка,
2. FPR – False Positive Rate или доля ошибочно предсказанных дефолтов,
3. TPR – True Positive Rate или доля правильно предсказанных дефолтов,
4. AUC – площадь под ROC кривой;

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах; пер. с англ. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.
2. Введение в статистическое обучение с примерами на языке R / Г. Джеймс [и др.]; пер с англ. С.Э. Мастицкого. – М.: МДК Пресс, 2017. – 456 с.
3. Hosmer, D.W. Applied Logistic Regression / D.W. Hosmer, J.S. Lemeshow // University of Massachusetts. – 2-nd ed. – 2000. – 376 p.
4. Микелуччи У. Прикладное глубокое обучение. Подход к пониманию глубоких нейронных сетей на основе метода кейсов / У. Микелуччи; пер. с англ. – СПб: БХВ-Петербург, 2020. – 368 с.
5. Грас Дж. Data Science. Наука о данных с нуля / Дж. Грас; Пер. с англ. – СПб.: БХВ-Петербург, 2017. – 336с.
6. Финансовая отчетность в соответствии с НСФО/ Белинвестбанк [Электронный ресурс]. – Режим доступа: https://belinvestbank.by/about-bank/fina nce-statistic/nsfo/. – Дата доступа: 06.03.2022.
7. Pandas, main page [Electronic resource]. – 02.04.2022. – Mode of access: <https://pandas.pydata.org/>. – Date of access: 10.04.2022.
8. Numpy, main page [Electronic resource]. – 2022. – Mode of access: <https://numpy.org/>. – Date of access: 10.04.2022.
9. Scipy, main page [Electronic resource]. – 05.02.2022. – Mode of access: <https://scipy.org/>. – Date of access: 10.04.2022.
10. Scikit-learn, machine learning in python [Electronic resuorce]. – 2021. – Mode of access: <https://scikit-learn.org/stable/>. – Date of access: 10.04.2022.
11. PyTorch, main page [Electronic resource]. – Mode of access: <https://pytorch.org/>. – Date of access: 10.04.2022.
12. Виды экономической деятельности, общегосударственный классификатор Республики Беларусь. – Минск, 2011. ­– 364 с.
13. Выброс (статистика) / Википедия, свободная энциклопедия [Электронный ресурс]. – 2021. – Режим доступа: <https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D1%80%D0%BE%D1%81_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)>. – Дата доступа: 10.04.2022.
14. Lopes, R.H.C. The two-dimentional Kolmogorov-Smirnov test / R.H.C. Lopes, I. Reid, P.R. Hobson // XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research. – 2007. – Amsterdam, Netherlands.
15. Loss functions/ PYTORCH DOCUMENTATION [Electronic resource]. – Mode of access: <https://pytorch.org/docs/stable/nn.html#loss-functions>. – Date of access: 16.04.2022.
16. TORCH.OPTIM/ PYTORCH DOCUMENTATION [Electronic resource]. – Mode of access: https://pytorch.org/docs/stable/optim.html. – Date of access: 16.04.2022.
17. Пойнтер, Я. Программируем с PyTorch: Создание приложений глубокого обучения / Я. Пойнтер. – СПб.: Питер, 2020. – 256 с.
18. Goboy, D. Understanding binary cross-entropy/ log loss: a visual explanation [Electronic resource]. – Mode of access: <https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a>. – Date of access: 23.04.2022.
19. Хендрик, Б. Машинное обучение / Б. Хендрик, Дж. Ричардс, М. Феверолф. – СПб.: Питер, 2017. – 336 с.
20. Тарик, Р. Создаем нейронную сеть.: Пер. с англ / Р. Тарик. – СПб.: ООО «Альфа-книга», 2017. – 272 с.
21. Novak, G. Building a One Hot Ecoding Layer with TensorFrow[Electronic resource] / Towards Data Science. – 07.06.2020. – Mode of access: https://towardsdatascience.com/building-a-one-hot-encoding-layer-with-tensorflow-f907d686bf39. – Date of access: 20.04.2022.
22. Шолле, Ф. Глубокое обучение на python / Ф. Шолее. – СПб.: Питер, 2018. – 400 с.