МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ПРИМЕНЕНИЕ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ АНАЛИЗА РИСКОВ МЕДИЦИНСКОГО СТРАХОВАНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Бутузова Владислава Андреевича

Научный руководитель
доцент, к. ф.-м. н., доцент _____ Л. В. Борисова

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2026

ВВЕДЕНИЕ

Актуальность данной работы обусловлена растущей потребностью страховых компаний в точных и надёжных алгоритмах оценки рисков для развития рынка добровольного медицинского страхования. Использование методов машинного обучения позволяет формировать более гибкие тарифы и минимизировать убытки страховых компаний.

Целью данной работы является исследование и оценка применимости методов машинного обучения для выявления закономерностей и прогнозирования величины рисков в медицинском страховании.

Для достижения поставленной цели необходимо решить следующие задачи:

— Провести первичный анализ набора данных.
— Выполнить предобработку данных.
— Обучить модель линейной регрессии.
— Разработать наиболее подходящую архитектуру и обучить полносвязную искусственную нейронную сеть.
— Провести сравнительный анализ полученных моделей.

Работа состоит из трёх основных разделов. В первом разделе рассматривается этап подготовки данных: производится выбор набора данных, проводится его предварительный анализ и осуществляется предобработка. Во втором разделе исследуется применение модели линейной регрессии для решения поставленной задачи. В третьем разделе описывается обучение модели нейронной сети: сначала излагаются основные теоретические сведения о многоуровневых перцептронах, функциях активации и методах оптимизации, затем реализуется процесс обучения, а также проводится анализ полученных результатов.

1 Работа с данными

1.1 Выбор набора данных

В машинном обучении для построения моделей могут использоваться данные самого разного формата — изображения, текст, аудиосигналы, временные ряды и другие. Однако в рамках данной работы рассматриваются табличные данные, которые являются одним из наиболее распространённых и понятных типов представления информации.

В табличных данных каждая строка соответствует отдельному объекту наблюдения или же записи, а каждый столбец содержит значения признаков, описывающих этот объект. Как правило, задача машинного обучения в этом формате заключается в предсказании одного из признаков — так называемой целевой переменной — на основе всех остальных признаков, выступающих в роли входных данных.

Эффективность применения методов машинного обучения во многом определяется качеством и особенностями используемого набора данных. В связи с этим первоочередным этапом работы является выбор и предварительный анализ подходящего датасета.

Для данной работы исходный набор данных был взят из репозитория Packt Publishing - британского издательства, специализирующегося на технической литературе и практических руководствах для IT-специалистов. Данные были импортированы с использованием функционала платформы Kaggle, представляющей собой онлайн-сообщество специалистов по Data Science и облачную среду для проведения соревнований по машинному обучению.

Выбранный набор данных основан на демографической статистике бюро переписи населения США. Он содержит такие параметры, как возраст, пол, ИМТ, количество детей, статус курения, регион и сумма страховой выплаты. В качестве целевой переменной будет рассматриваться признак суммы страховой выплаты, так как он отражает уровень риска страхования.

Риск страхования — это вероятность наступления страхового случая, связанного с убытками или расходами, которые страховая компания обязуется возместить согласно договору.

Было рассмотрено множество наборов данных, однако значительная их часть не показала удовлетворительных результатов при предварительной

проверке. Данный датасет был выбран по причине его доступности, релевантности теме исследования, а также достаточного объёма и структуры, позволяющих применять методы машинного обучения для анализа рисков страхования.

1.2 Анализ и предобработка данных

Перед началом непосредственной работы с любым набором данных необходимо провести его тщательный предварительный анализ и подготовку. Это позволяет заметить и устранить аномалии в выборке, а также выявить взаимосвязи между признаками до начала обучения, что упрощает задачу алгоритмам машинного обучения.

Для решения данной задачи, а также выполнения всех последующих операций с набором данных, была использована библиотека Pandas, предназначенная для удобной обработки и анализа табличных данных.

Целесообразно проверять данные на наличие дубликатов и удалять их, так как повторяющиеся записи создают ложный перекос в сторону определенных наблюдений, что ведет к переобучению модели на копиях. В рассматриваемом датасете обнаруженные дубликаты были удалены.

Необходимо выявлять пропуски в данных и обрабатывать их, так как записи с ними невозможно использовать в рассчётах. Такие записи либо удаляют, либо заполняют расчетными значениями. В исходном наборе данных пропусков обнаружено не было.

Значения, значительно отклоняющиеся от основной массы данных, называют выбросами. Они могут негативно отражаться на результатах модели. На выбросах модель получает огромный штраф во время обучения, из-за чего излишне подстраивается под эти значения. От чего точность для большинства примеров снижается. Чтобы заметить выбросы, можно анализировать распределения признаков на различных графиках.

Для большинства методов машинного обучения предпочтительным является нормальное распределение признаков, так как оно обеспечивает стабильность работы алгоритмов и более быструю сходимость градиентных ме-

тодов. Плотность нормального распределения описывается формулой:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{1}$$

Напротив, логарифмически нормальное распределение часто оказывается нежелательным при обучении моделей. Оно характеризуется выраженной асимметрией, что провоцирует проблему «естественных» выбросов. Формула плотности логарифмического распределения для $x > 0$ имеет вид:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}. \tag{2}$$

Для детального анализа структуры данных и оценки их близости к теоретическим распределениям в работе используются методы визуализации: построение гистограмм и ядерная оценка плотности (Kernel Density Estimation, KDE). Метод KDE позволяет получить сглаженную непрерывную кривую плотности распределения на основе имеющейся выборки:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{3}$$

где K — функция ядра, а h — сглаживающий параметр. В данной работе в качестве функции ядра используется ядро Гаусса:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}. \tag{4}$$

На рисунке 1 представлен график плотности распределения страховых выплат. Анализ KDE-кривой и столбцов гистограммы показывает, что подавляющее большинство наблюдений сосредоточено в области малых значений, однако в правой части графика наблюдается протяженная область редких, но экстремально высоких выплат, значительно удалённых от среднего. Такое явление называют тяжёлым хвостом. Визуально структура данных соответствует логнормальному распределению.

Такая форма распределения является типичной для актуарных данных и сферы страхования в целом. Подобные данные создают проблему, схожую с

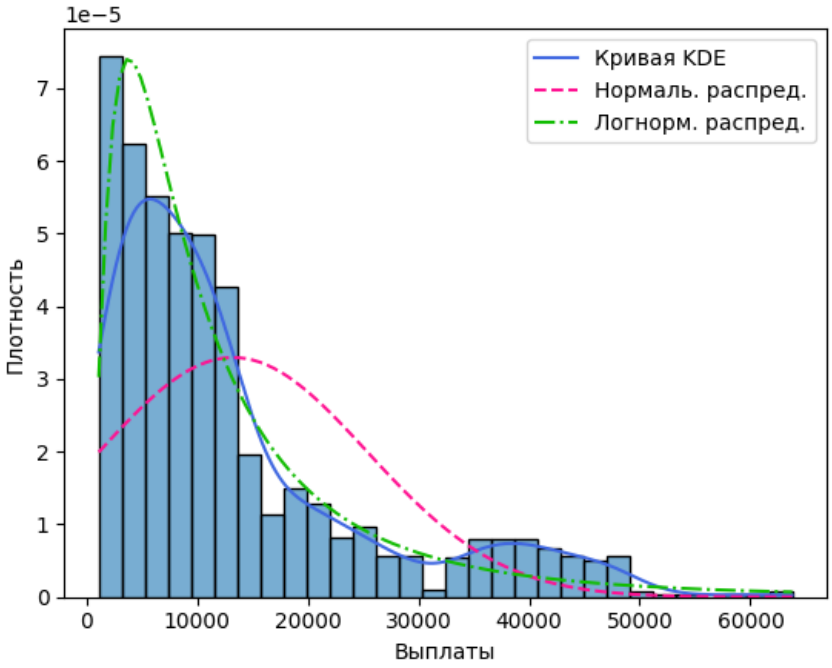

Рисунок 1 – График плотности распределения выплат

влиянием аномальных выбросов. Для минимизации описанного негативного влияния и приведения распределения целевой переменной к виду, близкому к нормальному, было применено логарифмическое преобразование.

В рамках проектирования признаков была введена новая бинарная переменная, фиксирующая одновременное наличие у субъекта факторов курения и ожирения. Создание такого производного признака позволяет модели учитывать комбинированный эффект данных факторов риска.

Для обеспечения возможности математической обработки к категориальным признакам было применено бинарное кодирование. Данный метод преобразует каждую категорию в отдельный столбец с признаками 0 или 1, что позволяет алгоритмам машинного обучения корректно интерпретировать номинальные данные без внесения ложного иерархического порядка между ними.

2 Линейная регрессия

2.1 Теория линейной регрессии

Линейная регрессия представляет собой метод моделирования связи между зависимой переменной y (целевым признаком) и набором независимых переменных $\mathbf{X}$ (признаков). В векторном виде математическая модель выражается следующим образом:

$$\hat{y} = \mathbf{X}\mathbf{w} + b, \tag{5}$$

где $\hat{y}$ — предсказанное значение, $\mathbf{X}$ — матрица признаков, $\mathbf{w}$ — вектор искомых весов (коэффициентов), а b — свободный член (смещение).

Целью обучения модели является нахождение таких параметров $\mathbf{w}$ и b, которые минимизируют функционал ошибки. В данной работе в качестве функции потерь используется среднеквадратичная ошибка (Mean Squared Error, MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \to \min_{\mathbf{w},b}, \tag{6}$$

где y_i — истинные значения целевого признака, а n — общее количество наблюдений.

Для нахождения оптимальных весов $\mathbf{w}$ в аналитическом виде используется метод наименьших квадратов (МНК). Если объединить смещение b с вектором весов $\mathbf{w}$ (добавив единичный столбец к матрице $\mathbf{X}$), решение системы нормальных уравнений имеет вид:

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{7}$$

Также для минимизации функционала ошибки в задачах с большим объемом данных могут применяться численные методы, такие как градиентный спуск.

Для оценки качества полученной модели используются следующие метрики:

1. Коэффициент детерминации R^2, описывающий долю объясненной дисперсии:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{8}$$

2. Средняя абсолютная процентная ошибка (MAPE), позволяющая оценить точность прогноза в относительных величинах:

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \tag{9}$$

2.2 Результаты обучения

В ходе оценки качества построенной модели были получены следующие значения метрик:

Метрика	**Значение**
Коэффициент детерминации (R^2)	0.52
Средняя относительная ошибка (MAPE)	26%

Анализируя полученные данные, можно сделать вывод, что результат является не вполне удовлетворительным. Значение коэффициента детерминации $R^2 = 0.44$ указывает на то, что модель объясняет лишь 44% дисперсии целевой переменной, что свидетельствует о недостаточной предсказательной способности в рамках текущего набора признаков.

3 Искусственные нейронные сети

3.1 Теория нейронных сетей

Искусственные нейронные сети (ИНС) представляют собой математические модели, построенные по принципу биологических нейронных сетей. Важнейшим компонентом ИНС являются функции активации, которые вносят в модель нелинейность, позволяя сети аппроксимировать сложные зависимости, отличные от линейных. В данной работе рассматриваются две функции активации:

1. Гиперболический тангенс (tanh):

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{10}$$

2. ReLU (Rectified Linear Unit), отсекающая отрицательные значения:

$$\mathrm{ReLU}(x) = \max(0, x). \tag{11}$$

3. Сигмоида, сжимающая входное значение в диапазон $(0, 1)$:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{12}$$

Процесс обучения нейронной сети заключается в итеративном обновлении параметров (весов и смещений), которые изначально инициализируются случайными значениями. Для минимизации функции потерь используется метод градиентного спуска, при котором изменение параметров происходит в направлении, противоположном вектору градиента:

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \eta \cdot \nabla_{\boldsymbol{\theta}} L, \tag{13}$$

где $\boldsymbol{\theta}$ — вектор всех параметров, η — скорость обучения (learning rate), а $\nabla_{\boldsymbol{\theta}} L$ — градиент функции потерь.

Для более эффективной оптимизации в работе применяется алгоритм Adam (Adaptive Moment Estimation). Он сочетает в себе идеи накопления инерции (momentum) и адаптивной корректировки скорости обучения для каждого параметра. Математический аппарат обновления весов в Adam опи-

сывается следующей системой уравнений:

$$\mathbf{g}_t = \nabla_{\boldsymbol{\theta}} L_t(\boldsymbol{\theta}_{t-1}); \tag{14}$$

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t; \tag{15}$$

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2; \tag{16}$$

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1-\beta_1^t}; \quad \hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1-\beta_2^t}; \tag{17}$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta \cdot \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \varepsilon}. \tag{18}$$

Здесь $\mathbf{m}_t$ и $\mathbf{v}_t$ — оценки первого и второго моментов градиента соответственно, а β_1, β_2 — гиперпараметры затухания. Данный оптимизатор позволяет ускорить сходимость и стабилизировать процесс обучения на сложных ландшафтах функции потерь.

3.2 Результаты обучения

Науилучшие результыта показало обучение модели с использованием оптимизатора Adam и нормализацией входных данных с помощью StandardScaler:

$$x' = \frac{x - \mu}{\sigma}.$$

В ходе оценки качества были получены следующие значения метрик:

Метрика	**Значение**
Коэффициент детерминации (R^2)	0.87
Средняя относительная ошибка (MAPE)	15%

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы были получены следующие значимые результаты. На этапе предобработки данных применение логарифмического преобразования целевой переменной позволило эффективно минимизировать влияние выбросов, что обеспечило статистическую стабильность процесса обучения моделей. Сравнительный анализ различных архитектур показал, что разработанная многослойная нейронная сеть превосходит классическую линейную регрессию, тем самым подтверждая наличие сложных нелинейных зависимостей в актуарных данных. Выбор оптимизатора Adam позволил достичь наиболее рационального баланса между скоростью сходимости алгоритма и точностью аппроксимации целевого функционала.

Практическая значимость работы заключается в возможности автоматизации процессов оценки страховых рисков в секторе добровольного медицинского страхования. Предложенный алгоритм обладает потенциалом к масштабированию и может быть интегрирован в информационные системы страховых компаний для динамического формирования тарифов и повышения точности резервирования.