МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ

(государственный университет)

ФАКУЛЬТЕТ УПРАВЛЕНИЯ И ПРИКЛАДНОЙ МАТЕМАТИКИ

Базовая организация - ФИЦ ИУ РАН

Кафедра «Интеллектуальные системы»

специализация «Интеллектуальный анализ данных (ИАД)»

Квалификационная работа на соискание степени магистра

по направлению 03.04.01 «Прикладные математика и физика»,

магистерская программа 010956 «Математические и информационные технологии»

Построение интерпретируемых моделей глубокого обучения в задаче социального ранжирования

Студент группы 274а

Гончаров Алексей Владимирович

Научный руководитель:

д.ф.-м.н.

Стрижов Вадим Викторович

Москва, 2018

Содержание

1 Аннотация

Решение задачи моделирования кредитного риска заемщика важно для основной деятельности финансовых организаций. На основе результата работы модели выносится решение о предоставлении кредитного заема физическому или юридическому лицу. В работе решается задача построения скоринговой карты и модели кредитного скоринга на ее основе. Из-за внутрибанковских требований от руководства отделов аналитики и внешних регуляторов необходимо рассмотреть в качестве моделей классификации класс интерпретируемых моделей, в частности, модель логистической регрессии. Такое ограничение связано с юридической чистотой проводимых в банке процессов и защищенностью банка от судебных исков со стороны клиентов. Целью работы является преобразование и расширение исходного признакового пространства с целью повышения качества решения задачи при минимальном повышении сложности модели классификации. Особенностью исследования является построение решения как суперпозиции в виде модели глубокого обучения, позволяющей включить необходимые процедуры порождения признаков в единую модель. Основная идея работы — построение единой процедуры оптимизации последовательной суперпозиции с использованием различных интерпретируемых модулей обработки признаков и построения моделей. В работе в одной суперпозиции модели совмещены процедуры сегментации, группировки и порождения признаков. Предлагается для каждой из процедур заменить алгоритм оптимизации локальных критериев качества оптимизацией пространства параметров суперпозиции в единой итерационной процедуре оптимизации. Для достижения наилучшего качества решения с точки зрения банковских критериев в процедуре оптимизации используется целевой эксплуатационного критерия качества, AUC или площадь под ROC-кривой. Качество решения задачи предложенным методом оценивается на открытых данных задач социального ранжирования и сравнивается с базовым решением и более сложными неинтерпретируемыми моделями.

Ключевые слова: социальное ранжирование; скоринговая карта; кредитный скоринг; глубокое обучение; суперпозиция моделей; порождение признаков; оптимизация параметров.

2 Введение

Актуальность темы. Решение задачи кредитного скоринга определяется как моделирование, использующееся для оценки риска кредитного заемщика или сущетвующих заемов в кредитном портфеле. Цель решения задачи кредитного сокринга — вынесение решения о предоставлении заемщику кредита на основе оценки вероятности дефолта.

Для финансовых учреждений, банков потребительского кредитования и розничной торговли, решение задачи кредитного скоринга является основой деятельности. В задаче построения моделей кредитного скоринга определяется уровень платежеспособности заемщика по его личной информации.

Аналитические отделы крупных банковских структур работают над построением скоринговых карт, содержащих неободимую для принятия решения о выдаче кредита информацию. Так как использование моделей, основанных на таких картах является основой бизнеса банковских структур, то разработка оценочных карт — ключевая компетенция управления рисками розничного банка при оценке кредитоспособности отдельного лица.

Аналитик не способен вручную выявить большое число закономерностей, присутствующих в данных. В такой сложной для аналитика задаче используются технологии автоматизации процесса выставления баллов клиентам, совмещаются различные скоринговые и риск модели.

Производительность моделей очень важна, и улучшение точности прогноза может привести к значительной экономии ресурсов розничного банка. Поэтому анализ различных моделей и алгоритмов разработки оценочных карт является актуальным для кредитных банков.

Для руководителей подобных организаций важна обоснованность проставленного скорингового балла во избежании скандальных ситуаций вокруг компании. Это приводит к использованию внутри банковских структур интерпретируемых моделей, которые зачастую имеют более низкое качество по сравнению с неинтерпретируемыми аналогами, построенными на основе нейронных сетей или неинтерпретируемых ансамблях моделей. Также накладываются внешние ограничения от управдяющих органов [6] на зависимости внутри скоринговых карт, которым должны удовлетворять построенные модели, что и приводит к требованию их интерпретируемости.

Активно развиваются два подхода к построению скоринговых карт и моделей. Современный подход использует новые неинтерпретируемые технологии, позволяющие получать высокое качество [8–11]: нейронные сети, бустинг над решающими деревьями. Аналитики крупных банков используют перечисленные модели для подтверждения или опровержения основных результатов, полученных с помощью интерпретируемых моделей. В случае несогласованности результатов интерпретируемых и неинтерпретируемых моделей запрос на предоставление кредита посылают на перепроверку аналитику.

Общепринятый в банках подход основан на использовании интерпретируемых ранжирующих и классифицирующих моделях, таких как логистическая регрессия [12–16]. В литературе предлагаются методы порождения признаковых пространств, и модификации этих методов. Полученные

модели позволяют выделить необходимые интерпретируемые признаки, которые будут одобрены аналитиками организации.

Для определения уровня кредитоспособности заемщика используется анкета заемщика, содержащая признаки как в линейных, так и в ординальных и номинальных шкалах. Такие анкеты строятся аналитиками, которые проводят социальные исследования по влиянию различных факторов на кредитоспособость. В анкету включаются наиболее распространенные у заемщиков признаки, проверка которых осуществляется просто. Последующая обработка скоринговых карт включает все процедуры, которые аналитик выполняет при построении эксплуатируемой модели. А именно: процедуры группировки [17] и сегментации, создание интерпретируемых комбинаций признаков, построение моделей классификации и вычисление критерия качества. Базовый принцип построения модели — это суперпозиция процедур порождения признаков.

Так в работе [4] предлагается использовать биннинг или сегментацию как один из этапов обработки признакового пространства. Сегментация отображает непрерывный набор признаков в небольшие группы или категории. Этот классический шаг построения скоринговой карты [5] аналитики признают устойчивым и интерпретируемым решением проблемы представления скрытых нелинейностей. Настоящая работа использует подход, описанный в [4] как базовый вид процедуры сегментации предлагаемой модели. В [2] для обработки данных малого объема и создания кредитных карт на их основе предлагается использование метрических моделей классификации. Работа [7] предлагает проводить кластеризацию исходной выборки для последующего построения моделей для каждого отдельного кластера с целью повышения качества без потери интерпретируемости модели. Такой подход позволяет в том числе решить проблему наличия разнородностей в данных. Настоящая работа проводит исследование о преобразовании признакового пространства, поэтому не используется методы разделения прсотранства объектов.

Актуальность настоящей работы обоснована нуждами финансовых организаций по автоматизированному построению скоринговых карт и моделей и по повышению качества их работы без существенного повышения сложности и без потери интерпретируемости модели. Также требуется построить интерпретируемую модель в целом и каждую из процедур для обоснованности применения таких моделей. В настоящий момент времени отсутствует единая процедура построения скоринговых моделей, позволяющая объединить в единую суперпозицию.

В задачах кредитного скоринга используют два вида критериев качества: оптимизационные и эксплутационные. Эксплуатационные критерии качества заданы исходя из потребностей индустрии при решении бизнес-задач. Оптимизационные критерии влияют на настройку параметров модели и нужны для ее построения, но не связаны с эксплуатационными критериями.
Эксплуатационные критерии качества:

- AUC — площадь под ROC-кривой (4). Критерий качества в задачах бинарной классификации, имеющий физический смысл доли правильно упорядоченных алгоритмом пар объектов. Критерий используется для оптимизации модели и для ее оценки экспертами. Это основной критерий сравнения моделей.

- Структурная сложность, число признаков. От сложности модели напрямую зависит ее интерпретируемость экспертами, которые выставляют соответствующие требования. В предложенной модели этот критерий зависит от значения структурных параметров модели, поэтому в оптимизируемый функционал 4 значение не входит.

- Неотрицательность весов модели. Требование неотрицательности весов позволяет проводить анализ используемых признаков модели.

- Устойчивость во времени. Такое ограничение вызвано высокими рисками финансовых организаций при нестабильном поведении модели.

Оптимизационные критерии качества:

- Правдоподобие модели. Правдоподобие модели указывает на вероятностный смысл полученного ответа и используется в построении решения для поиска оптимальных параметров логистической регрессии.

- Устойчивость — ортогональность признаков. Для устойчивости модели необходимо отсутствие мультиколлинеарности признаков. Чем разнороднее они будут, тем лучше будет вести себя модель.

- Точность, полнота. В некоторых бизнес-приложениях точность и полнота метода позволяет вычислять финансовые риски используемой модели.

Для исследования применимости предложенного решения и его сравнения с базовыми и альтернативными подходами предлагается использовать эксплуатационный критерий качества в единой постановке задачи оптимизации. В [3] проводится обзор методов сравнения скоринговых карт и предлагается комплексная методика их оценивания. AUC или площадь по ROC-кривой — самый распространенный критерий качества оценки скоринговых моделей. В работе [1] предлагается использовать AUC как целевой и единственный критерий качества в задаче оптимизации. Такой же подход используется в настоящей работе.

Цель работы. Целью работы является создание модели глубокого обучения для построения скоринговой карты как суперпозиции интерпретируемых процедур порождения признаков. Используется принцип глубокого обучения, который подразумевает оптимизацию целевого значения функции ошибки внутри каждой отдельной процедуры. Необходимо при этом удовлетворить требованиям интерпретируемости и качества итоговой модели при заданной невысокой сложности.

Научная новизна. каждый из которых доставляет максимум локальному критерию качества и не зависит от других. В работе предполагается, что такая постановка задачи ведет к поиску не оптимального решения задачи.

Предложена новая процедура построения интерпретируемых скоринговых моделей методами глубокого обучения. Проблема неоптимальности решения из-за оптимизации локальных критериев качества решается путем использования принципа глубокого обучения: оптимизации единого глобального критерия качества внутри каждого из шагов.

Также возникает проблема неоптимальности из-за отсутствия итерационной процедуры оптимизации каждого шага, которая будет учитывать изменения параметров других шагов порождения признаков. Предлагается реализация итеративного алгоритма оптимизации всей суперпозиции. Для этого изменены процедуры оптимизации сегментации, группировки и порождения признаков.

Практическая ценность. Результаты данной работы можно использовать для преобразования признакового пространства в задачах анализа данных, решать задачи бинарной классификации и социального ранжирования. Резльтаты применимы для банковских организаций, выявление интерпретируемых социальных групп, определение суперпозиций признаков, объединение категорий в интерпретируемые группы.

Положения, выносимые на защиту:

- Исследована проблема порождения интерпретируемых моделей в виде суперпозиции модулей
- Предложен метод единой оптимизации параметров суперпозиции, решающий указанную проблему
- Разработан алгоритм построения модели социального ранжирования согласно предложенному методу
- Решена задача кредитного скоринга на открытых данных при помощи разработанного алгоритма

3 Постановка задачи социального ранжирования

Дана выборка $D = \{(x_i, y_i) : i = 1, \ldots, m\}$, содержащая m объектов. Каждый объект описан линейными, категориальными и бинарными признаками, $x_i \in \mathbb{R}^l \times \mathbb{C}^c \times \mathbb{B}^b$ и принадлежит одному из двух классов: $y_i \in \{0, 1\}$. Индексы объектов $\{i = 1, \ldots, m\} = \mathscr{I}$ выборки поделены $\mathscr{I} = \mathscr{L} \bigsqcup \mathscr{T}$ на индексы объектов обучения и контроля.

Обозначим матрицу плана $X \in \mathbb{R}^{m \times n}, X = (\chi^1, \ldots, \chi^n)$, где χ_j — j-ый признак. Зададим множество индексов всех признаков $\mathscr{A} = \mathscr{A}_l \sqcup \mathscr{A}_c \sqcup \mathscr{A}_b$, где $\mathscr{A}_l$ обозначает индесы признаков в линейных шкалах, $\mathscr{A}_c$ — категориальных признаков, а $\mathscr{A}_b$ — бинарных признаков.

Модель классификации F в работе представлена в виде суперпозиции $F = f \circ f_f \circ f_g \circ f_s$, где f_s производит сегментацию линейных признаков, f_g — группировку категорий для категориальных признаков, f_f — строит новые признаки на основе старых, причем все эти функции переводят объект x из одного признакового пространства в другое. Это вызывает изменение матрицы плана X. Суперпозиция $f_f \circ f_g \circ f_s$ отображает одну матрицу плана в другую:

$$f_f \circ f_g \circ f_s : X \mapsto \hat{X}, \tag{1}$$

где $\hat{X} = (\hat{\chi}^1, \ldots, \hat{\chi}^{\hat{n}})$.

В работе функцией f выбрана логистическая регрессия как интерпретируемая модель классификации над получившимся признаковым пространством:

$$f(\hat{X}, w) = \frac{1}{1 + \exp(-\hat{X}w)}. \tag{2}$$

В качестве целевого критерия качества $L(w)$ для обучения модели логистической регрессии используем логарифмическую функцию правдоподобия:

$$L(w) = -\ln P(D|w) = -\sum_{i \in \mathscr{L}} \big(y_i \ln w^{\mathsf{T}} x_i + (1 - y_i) \ln(1 - w^{\mathsf{T}} x_i)\big). \tag{3}$$

При оптимизации параметров процедур сегментации, группшировки и порождения признаков, а также для оценки качества эксплуатационным критерием, согласно банковским требованиям, используется площадь под ROC-кривой:

$$Q(w) = \frac{\sum_{i \in \mathscr{L}} \sum_{j \in \mathscr{L}} [y_i < y_j][F(x_i) < F(x_j)]}{\sum_{i \in \mathscr{L}} \sum_{j \in \mathscr{L}} [y_i < y_j]}. \tag{4}$$

Функция f является параметрической с параметрами w. Для f_f, f_g и f_s помимо параметров $P = p_f \sqcup p_g \sqcup p_s$ введены структурные параметры $H = h_f \sqcup h_g \sqcup h_s$. Задача подбора оптимальных параметров w, P и структурных параметров H модели классификации F сводится к решению задачи оптимизации:

$$H^*, P^*, w^* = \operatorname*{argmin}_{H,P,w} Q(w). \tag{5}$$

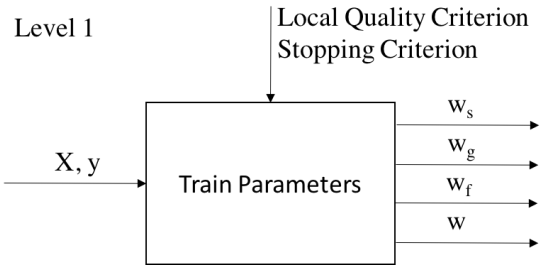

Рис. 1: Структурная диаграмма базового подхода, верхний уровень диаграммы

4 Задачи оптимизации параметров и структуры модели

Базовый подход к решению оптимизационной задачи 5 предполагает оптимизацию локальных критериев качества для каждой из процедур и изображен в виде структурной диаграммы на рисунках 1 и 2:

$$h_s^*, p_s^* = \underset{h_s, p_s}{\operatorname{argmin}} S_s(h_s, p_s), \tag{6}$$

$$h_g^*, p_g^* = \underset{h_g, p_g}{\operatorname{argmin}} S_g(h_g, p_g), \tag{7}$$

$$h_f^*, p_f^* = \underset{h_f, p_f}{\operatorname{argmin}} S_f(h_f, p_f), \tag{8}$$

$$w^* = \underset{w}{\operatorname{argmin}} L(w), \tag{9}$$

где S_s, S_g, S_f — локальные критерии качества для базового подхода, описаны далее.

Обучение модели F выполняется следующим образом. Дана обучающая выборка, для каждой функции из суперпозиции $F = f \circ f_f \circ f_g \circ f_s$ последовательно подбираются параметры, оптимальные для соответствующего критерия качества. Процедуры поиска параметров функций независимы. Существенное изменение параметров одной из функций не приводит к изменению параметров в других функциях. Такая независимость внутри единого пространства параметров всей модели классификации и оптимизация локальных критериев качества, которые не связаны с единым целевым эксплуатационным критерием, приводят к неоптимальности решения относительно целевого оптимизируемого критерия.

Утверждение 1. Для задачи оптимизации: $S(x_1, \ldots, x_n) \to \min$ при начальном приближении $X^0 = x_1^0, \ldots, x_n^0$ процедура оптимизации:

$$\hat{x}_i = \underset{x_i}{\operatorname{argmin}} S_i(\hat{x}_1, \ldots, \hat{x}_{i-1}, x_i, x_{i+1}^0, \ldots, x_n^0), i = 1, \ldots, n,$$

доставляет качество не лучше, чем процедура оптимизации:

$$\widetilde{x}_1, \ldots, \widetilde{x}_n = \underset{x_1, \ldots, x_n}{\operatorname{argmin}} S(x_1, \ldots, x_n),$$

то есть: $S(\hat{x}_1, \ldots, \hat{x}_n) \leq S(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

Доказательство: $\forall x_i \neq \widetilde{x}_i : S(\widetilde{x}_1, \ldots, x_i, \ldots, \widetilde{x}_n) \leq S(\widetilde{x}_1, \ldots, \widetilde{x}_i, \ldots, \widetilde{x}_n)$. $\square$

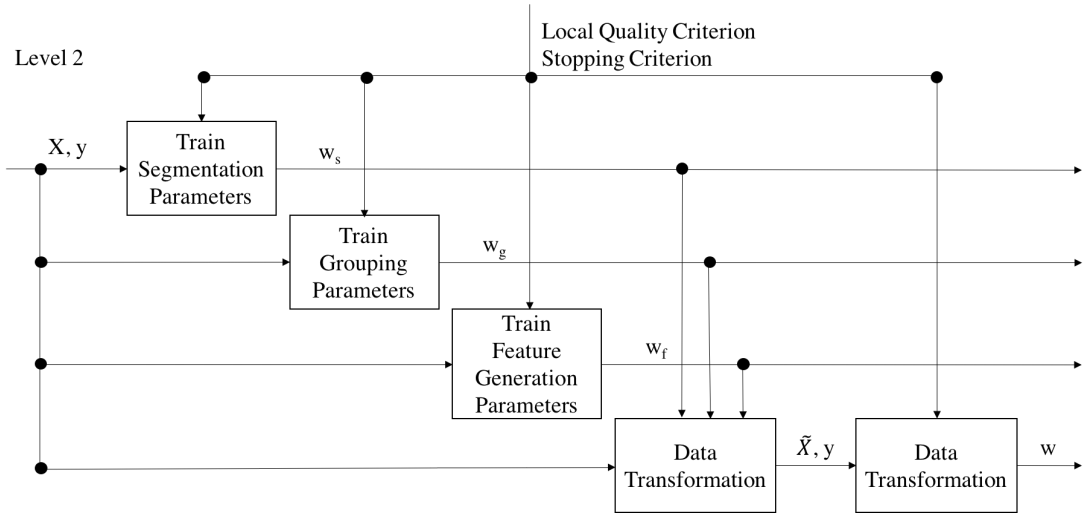

Рис. 2: Структурная диаграмма базового подхода, нижний уровень диаграммы

При решении задачи 5 поиска оптимальных параметров модели классификации оптимизация происходит по всему пространству параметров одновременно. В работе для этого предлагается в первую очередь использовать единый целевой критерий вместо локальных критериев качества, который и будет влиять на последовательное обучение каждого отдельного блока. Такая процедура обучения изображена в структурной диаграмме на рисунках 3 и 5. Изменение относительно базового подхода мотивировано поиском лучшего, нежели базового, решения задачи (5) согласно глобальному критерию качества.

При решении задачи (5) на итоговый критерий качества влияет каждая из оптимизируемых переменных H, P, w. Процедура оптимизациив базовом подходе проводится последовательно для каждой группы переменных. Из-за отсутствия итерационного алгоритма оптимизации внутри различных подпространств пространства параметров модели на каком-то из шагов оптимизации суперпозиции оптимальное решение перестает быть оптимальным. Предлагается создать итерационный алгоритм оптимизации суперпозиции, где начальным приближением параметров модели для очередной итерации является результат последовательной оптимизации суперпозиции на предыдущей итерации. Структурная диаграмма предложенного алгоритма изображена на рисунках 4 и 5.

Такой выбор предполагается оптимальным в задаче: пусть $\hat{x}_1 \sqcup \cdots \sqcup \hat{x}_n = x_1^0 \sqcup \cdots \sqcup x_n^0$. Опишем алгоритм решения задачи в виде псевдокода $A(N)$, где N — число итераций алгоритма:

$for \quad i = 1, \ldots, N:$

$$\hat{x}_1 = \underset{x_1}{\operatorname{argmin}}\, \text{step} \quad Q(\hat{x}_1 \sqcup \cdots \sqcup \hat{x}_n)$$

$\ldots$

$for \quad i = 1, \ldots, N:$

$$\hat{x}_n = \underset{x_n}{\operatorname{argmin}}\, \text{step} \quad Q(\hat{x}_1 \sqcup \cdots \sqcup \hat{x}_n).$$

Предположение 2. Процедура численной оптимизации для решения задачи: $Q(x_1, \ldots, x_n) \to \min$, представимая в виде псевдокода $A(N)$, не лучше, чем следующая процедура оптимизации, описанная в виде псевдокода: $for\ j = 1, \ldots, M : A(\frac{N}{M})$

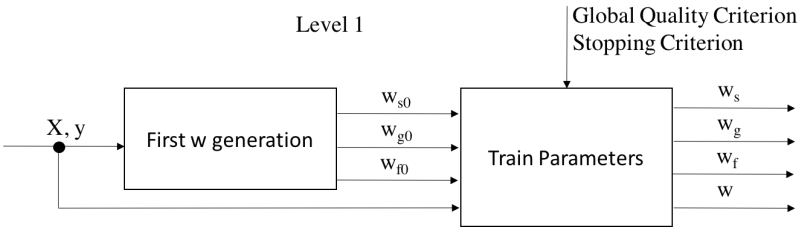

Рис. 3: Структурная диаграмма подхода с единым функционалом качества, верхний уровень диаграммы

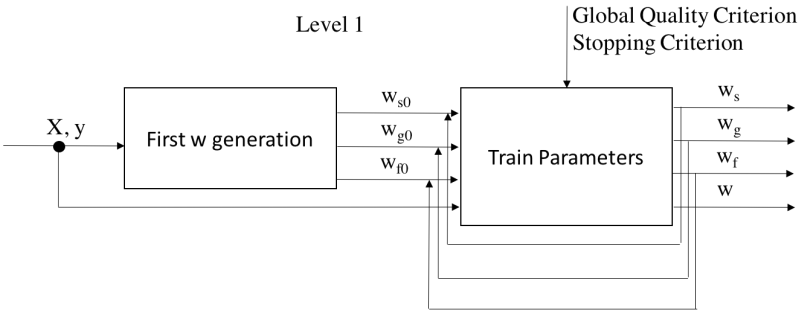

Рис. 4: Структурная диаграмма итерационного подхода с единым функционалом качества, верхний уровень диаграммы

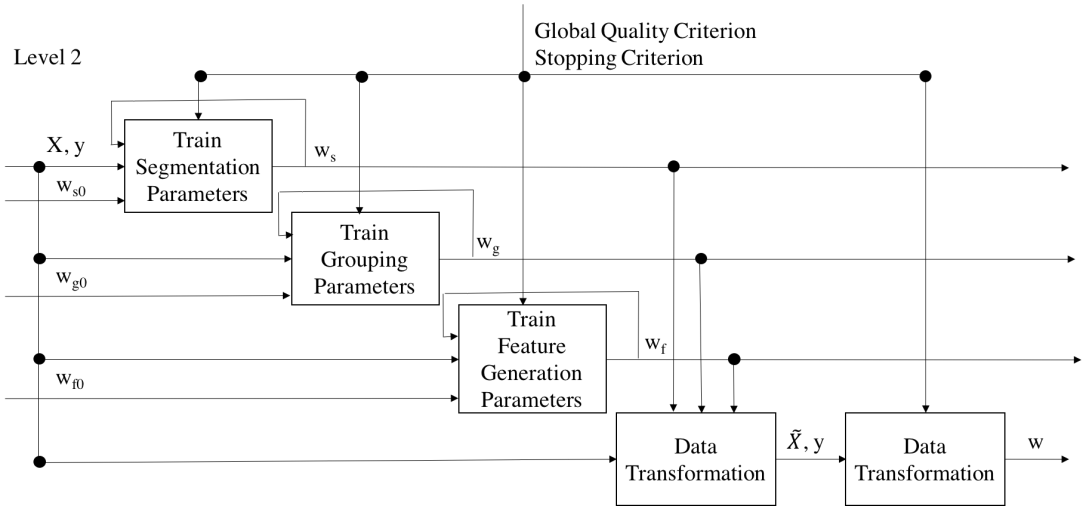

Рис. 5: Структурная диаграмма подхода с единым функционалом качества, верхний уровень диаграммы

Заметим, что при таком алгоритмическом решении задачи численной оптимизации на каждом цикле могут быть получены различные матрицы $\hat{X}$. Это приводит к неоднозначности выбора начальных параметров функции f при оптимизации цикла. Предлагается воспользоваться высокой скоростью сходимости алгоритма Ньютона-Рафсона оптимизации модели логистической регрессии и выполнять единую процедуру оптимизации модели f в каждом цикле.

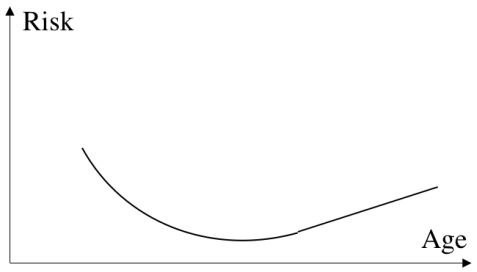

Рис. 6: Пример нелинейной зависимости уровня риска от возраста.

5 Сегментация линейных признаков

Сегментация дает возможность учета сложной кусочно-постоянной немонотонной зависимости функции ошибки от признака в монотонной линейной шкале и упростить интерпретацию итоговой модели. Исходное признаковое пространство дополняется бинарными признаками, обозначающими принадлежность исходного признака заданному сегменту на линейной шкале. Пример такой нелинейной зависимости уровня риска от возраста приведен на рисунке 6.

Предположение 3. Если $f(x)$ обладает высокой степенью нелинейности, то линейная модель

$$y_1 = w_1 x + w_2$$

не лучше, чем линейная модель, построенная на бинарных признаках, полученных в результате сегментации признака x:

$$y_2 = w_i, \quad x \in [a_i, b_i],$$

то есть: $||f(x) - y_2|| \leq ||f(x) - y_1||$.

5.1 Постановка задачи сегментации

Структурным параметром h_s процедуры сегментации является набор количества узлов разбиения для каждого из признаков. Пусть значения признака χ принадлежат отрезку $[a,b]$. Параметрами p_s процедуры сегментации является набор координат узлов разбиения, причем $p_s^j \in [a,b], j = 0,\ldots,h_s; \quad a = p_s^0 < \cdots < p_s^j < \cdots < p_s^{h_s} = b$.

Исходная матрица плана X пополняется новым набором бинарных признаков, являющихся индикаторами принадлежности значения исходного линейного признака каждому сегменту:

$$\{\chi^j\}_{j=1}^{h_s} \in \mathbb{B}^{h_s} : \chi^j = \left[\chi \in [p_s^{q-1}; p_s^q]\right], q = 1, ..., h_s.$$

Подбор структурных параметров и параметров процедуры сегментации набора линейных признаков с индексами $\mathscr{A}_l$ на обучающей подвыборке выборки D происходит путем решения следующей оптимизационной задачи для целевой функции:

$$h_s^*, p_s^*, w^* = \underset{h_s, p_s, w}{\operatorname{argmin}} Q(h_s, p_s, w | D, \mathscr{A}_l, \mathscr{L}),$$

где Q — критерий качества (4).

5.2 Решение задачи сегментации

Параметрами задачи сегментации служат узлы разбиения, принадлежащие действительной шкале. В базовой модели для поиска параметров и структурных параметров сегментов используются значения WOE выбранного сегмента q и доля выборки в сегменте q для признака i [?]:

$$\text{count}_q = \frac{1}{m} \sum_{j \in \mathscr{L}} \left[\chi_j \in [p_s^{q-1}; p_s^q] \right],$$

$$\text{woe}_q = \log \frac{g_j}{b_j} = \log \frac{\sum_{j \in \mathscr{L}} \left[\chi_j \in [p_s^{q-1}; p_s^q] \right] [y_j = 1]}{\sum_{j \in \mathscr{L}} \left[\chi_j \in [p_s^{q-1}; p_s^q] \right] [y_j = 0]},$$

где g_j и b_j — доли классов 0 и 1 в j-ом сегменте.

Для построения базовой модели решения (6) аналогично [4] предлагается делить интерпретируемый линейный признак на максимально возможное число сегментов, а затем объединять их согласно нескольким принципам: доля объектов в сегменте не менее 4% от общего числа выборки, значение woe для соседних сегментов существенно отличается. Таким образом локальный критерий качества S_s для признака χ: $S_s = [count_q \geq 0.04] \quad [|woe_q - woe_{q-1}| \geq 0.1]$.

Новизна настоящей работы состоит в использовании численных методов оптимизации для поиска параметров сегментации выбранного признака. Целевым критерием качества служит $Q(w)$, где w обозначает все параметры модели. Параметрами оптимизации для каждого признака χ являются p_s, значит целевой критерий качества $Q(p_s)$.

Итерация оптимизации параметров процедуры сегментации признака χ представляет собой градиентный спуск внутри пространства параметров модели:

$$p_s^i = p_s^{i-1} - \lambda \bigtriangledown Q(p_s),$$

где$\bigtriangledown Q(p_s)$ найден численными методами.

Итерация оптимизации параметров процедуры сегментации в суперпозиции модели — несколько циклов последовательных итераций градиентного спуска для каждого из признаков $\chi \in X$. После каждой итерации численной оптимизации производится процедура корректировки параметров сегментации

$$Correction(p_s): \quad \text{удаляется} \quad p_s^q \quad \text{, если} \quad p_s^q - p_s^{q-1} < 2 \tag{10}$$

Таким образом при консолидации параметров сегментации в одной точке линейного пространства структурный параметр сегментации признака χ изменяется. Структурные параметры будут найдены автоматически.

$for \quad iteration = 1, \ldots, N:$

$\quad for \quad \chi \in X:$

$\quad\quad p_s^i = p_s^{i-1} - \lambda \bigtriangledown Q(p_s)$

$\quad\quad p_s^i = Correction(p_s^i) \qquad (10)$

Определение. Сложностью процедуры сегментации для признака χ является число параметров или значение структурного параметра h_s.

Определение. Сложностью процедуры сегментации модели является число параметров сегментации модели или сумма значений структурного параметра $\sum_{\chi \in X} h_s$.

6 Группировка категориальных признаков

В скоринговых картах возникают категориальные признаки с большим числом категорий. Обработка таких неупорядоченных признаков сильно повышает сложность итоговой модели. Например, категориальный признак профессии заемщика при его кодировании может создать сотни бинарных признаков, сильно разреженных. Такое число бинарных признаков не только повышает сложность признакового пространства, но способствует переобучению сложных моделей, а также снижает интерпретируемость построенных решений. Для решения описываемой проблемы предлагается объединять категории признаков в группы, чтобы резко снизить сложность модели и признакового пространства без снижения качества итогового решения. Такой процесс называется группировкой нескольких категорий в одну, снижает сложность модели и повышает интерпретируемость.

6.1 Постановка задачи группировки

Пусть категориальный признак χ имеет $|C|$ категорий, где C — множество категорий этого признака. Структурным параметром h_g процедуры группировки признака χ является количество групп новых категорий, $|h_g| < |C|$. Параметром группировки является сюръекция $p_g : C \to h_g$. Исходное признаковое пространство изменяется: категориальный признак χ заменяется категориальным признаком χ_h.

$$\begin{array}{lcccccll} \chi & = & 1 & 2 & 3 & \dots & \text{C} & \text{C — число категорий} \\ & & \downarrow & \downarrow & \downarrow & & \downarrow & \\ \chi_h & = & \gamma_1 & \gamma_2 & \gamma_3 & \dots & \gamma_C & |h_g| \text{ — число новых категорий, } \gamma_i \in h_g \end{array}$$

Подбор структурных параметров и сюръекций процедуры группировки набора категориальных признаков с индексами $\mathscr{A}_c$ на обучающей подвыборке выборки D происходит путем решения следующей оптимизационной задачи для целевой функции:

$$h_g, p_g, w = \underset{h_g, p_g, w}{\operatorname{argmin}} Q(h_g, p_g, w | D, \mathscr{A}_c, \mathscr{L}),$$

где Q - критерий качества, описанный в (4).

6.2 Решение задачи группировки

Параметрами задачи сегментации служат сюръективные отображения. В базовой модели для поиска параметров и структурных параметров сегментов используются значения WOE выбранной категории c и доля выборки, принадлежащей этой категории для признака χ:

$$count_c = \frac{1}{m} \sum_{j \in \mathscr{L}} [\chi_j = c],$$

$$woe_c = \log \frac{g_j}{b_j} = \log \frac{\sum_{j \in \mathscr{L}} [\chi_j = c][y_j = 1]}{\sum_{j \in \mathscr{L}} [\chi_j = c][y_j = 0]},$$

где g_j и b_j — доли классов 0 и 1 в j-ом сегменте. Для построения базовой модели предлагается категории, для которых доля объектов мала объединять с ближайшими по значению WOE катего-

риями.

Для построения базовой модели решения (7) предлагается объединять близкие по значению woe_c категории в одну группу, а также категории, где доля объектов менее 4%, объединять с ближайшими по значению woe_c. Таким образом локальный критерий качества S_g для категориального признака χ: $S_g = [count_c \geq 0.04] \prod_{i,j}[|woe_i - woe_j| \geq 0.1]$.

В работе необходимо подобрать сюръекции, доставляющие оптимум критерию качества. Численные методы оптимизации не подходят для решения подобных задач. Новизна настоящей работы состоит в использовании методов оптимизации, основанных на принципах генетических алгоритмов, для поиска параметров группировки выбранного признака. Целевым критерием служит $Q(w)$, где w обозначает все параметры модели. Параметрами оптимизации для каждого категориального признака χ являются p_g, значит целевая функция $Q(p_g)$.

Определение. Особь a_χ в задаче оптимизации процедуры группировки категориального признака χ — последовательность $\{\gamma_1, \ldots, \gamma_C\}, \gamma_i \in h_g$, задающая произвольное сюръективное отображение p_g.

Определение. Процедура скрещивания $crossing(a^1_\chi, a^2_\chi)$ двух особей a^1_χ и a^2_χ — скрещивание соответствующих последовательностей: $\{\gamma^1_1, \ldots, \gamma^1_C\}$ и $\{\gamma^2_1, \ldots, \gamma^2_C\}$. Скрещивание двух таких последовательностей задается случайным бинарным вектором $bin \in \mathbb{B}^{|C|}$:

$a^1_{12} = \{bin_1\gamma^1_1 + (1-bin_1)\gamma^2_1, \ldots, bin_C\gamma^1_C + (1-bin_C)\gamma^2_C\}$.

$a^2_{12} = \{bin_1\gamma^2_1 + (1-bin_1)\gamma^1_1, \ldots, bin_C\gamma^2_C + (1-bin_C)\gamma^1_C\}$.

Итерация оптимизационного алгоритма в задаче группировки признака χ представляет собой процедуру заданного числа скрещиваний случайных особей из заданного набора последовательностей $A = \{a^i_\chi\}$, задающих сюръективные отображения и последующего отбора лучших особей из получившегося множества. Лучшая особь выбирается как очередное приближение p_g:

Итерация оптимизационного алгоритма в задаче группировки в суперпозиции модели — несколько циклов последовательных итераций генетического алгоритма для каждого из категориальных признаков $\chi \in X$. Так как при начальной инициализации и скрещивании особей нет ограничений на число новых категорий, то структурные параметры группировки будут найдены автоматически.

$for \quad iteration = 1, \ldots, N:$

$\quad for \quad \chi \in X:$

$\quad\quad for \quad j = 1, \ldots, N:$

$\quad\quad\quad A \quad append \quad crossing(a^i_\chi, a^k_\chi)$, где i,k — случайные

$\quad\quad\quad A \quad = \quad select \quad M \quad best \quad of \quad A$

$\quad\quad\quad p_g \quad = \quad select \quad one \quad best \quad of \quad A$

Определение. Сложностью процедуры группировки для признака χ является число категорий структурного параметра $|h_g|$.

Определение. Сложностью процедуры группировки является число новых категорий модели или сумма значений мощностей структурного параметра $\sum_\chi |h_g|$.

7 Порождение интерпретируемых признаков

Сложные модели доставляют наибольшее качество в решении задачи за счет выявления нелинейных закономерностей в данных и сложных взаимосвязей между признаками. В случае использования простых интерпретируемых моделей возможно учитывать такого вида взаимосвязи, если создавать на основе исходного признакового пространства новые признаки. Их можно конструировать как суперпозицию признаков и интерпретируемых операций. Должно быть задано множество операций на признаковом пространстве и правила порождения новых суперпозиций.

7.1 Постановка задачи порождения

Пусть множество $\{f_i\}$ — множество интерпретируемых порождающих непараметрических функций над признаками, заданных экспертами. Ниже приведена таблица с перечислением соответствующих операций и их свойств.

Описание	Формула	in	N in	out
Negate binary	$\bar{x}$	bin	1	bin
Logarithm	$\log(x)$	lin	1	lin
Logistic sigmoid	$\frac{1}{1+\exp(x)}$	lin	1	lin
Square root	$\sqrt{x}$	lin	1	lin
Inverse	$\frac{1}{x}$	lin	1	lin
Multiplication	$x*y$	any	2	lin
Sum	$x+y$	any	2	lin

Пусть $\{\chi_i\}, i \in \mathscr{A}_l \sqcup \mathscr{A}_b$ — измеряемые линейные и бинарные признаки. Необходимо найти лучшую комбинацию заданного числа из всех допустимых суперпозиций признаков $\{\chi_i\}$ с использованием функций $\{f_i\}$, после чего вычислить и добавить их в матрицу плана X.

Определение. Суперпозиция — формула, представимая в виде дерева, в каждом узле которого находится функция из $\{f_i\}$, а в каждом листе — признак из $\{\chi_i\}$. Сложность суперпозиции — число использованных в ней элементов из $\{f_i\}$ и $\{\chi_i\}$ с повторениями.

Из множества всех суперпозиций Σ необходимо выбрать наилучшую комбинацию суперпозиций $p_f = \{\mathbb{F}_i\}$. Структурными параметрами являются количество суперпозиций в комбинации $h_f = |\{\mathbb{F}_i\}|$ и максимальная сложность суперпозиции, а параметрами — комбинация $p_f = \{\mathbb{F}_i\}$. Подбор структурных параметров и параметров процедуры порождения на обучающей подвыборке выборки D происходит путем решения следующей оптимизационной задачи для целевой функции:

$$h_f, p_f, w = \underset{h_f, p_f, w}{\operatorname{argmin}} Q(h_f, p_f, w | D, \mathscr{A}_l \sqcup \mathscr{A}_b, \mathscr{L}),$$

где Q — критерий качества, описанный в (4).

7.2 Решение задачи порождения

Параметрами задачи порождения признаков служат наборы суперпозиций. В базовой модели для поиска параметров и структурных параметров генерации признаков используется корреляция каждой суперпозиции с целевой переменной:

$$corr_g(p_f) = \sum_{\mathbb{F}_i \in p_f} (1 - corr(\mathbb{F}_i, y)^2).$$

Для построения базовой модели (8) предлагается качество набора построенных признаков оценивать с помощью $corr_g$. Таким образом, локальный критерий качества S_f для поставленной задачи порождения новых признаков примет вид: $S_f = corr_g$

В работе необходимо подобрать набор суперпозиций, доставляющий оптимум критерию качества. Численные методы оптимизации не подходят для решения такой задачи. Новизна настоящей работы состоит в использовании методов оптимизации, основанных на принципах генетических алгоритмов, для поиска новых суперпозиций. Целевым критерием служит $Q(w)$, где w обозначает все параметры модели. Параметрами оптимизации являются p_f, значит целевой критерий $Q(p_f)$.

Определение. Особь a в задаче оптимизации процедуры порождения признаков — комбинация суперпозиций $\{\mathbb{F}_i\}$ или новых признаков модели.

Определение. Процедура скрещивания $crossing(a_1, a_2)$ двух особей a^1 и a^2 — скрещивание соответствующих комбинаций суперпозиций: $\{\mathbb{F}_i^1\}$ и $\{\mathbb{F}_i^2\}$. Скрещивание двух таких последовательностей задается случайным бинарным вектором $bin \in \mathbb{B}^{|\{\mathbb{F}_i\}|}$:

$a_{12}^1 = \{bin_i \mathbb{F}_i^1\} \cup \{(1 - bin_i)\mathbb{F}_i^2\}$.

$a_{12}^2 = \{bin_i \mathbb{F}_i^2\} \cup \{(1 - bin_i)\mathbb{F}_i^1\}$.

Итерация оптимизационного алгоритма в задаче порождения признаков в суперпозиции модели представляет собой несколько циклов заданного числа скрещиваний случайных особей $\{\mathbb{F}_i\}$ из заданного набора $F = \{\{\mathbb{F}_i\}_j\}$. Лучшая особь выбирается как очередное приближение p_g:

$for \quad iteration = 1, \ldots, N:$

$F \quad append \quad crossing(a^i, a^k)$, где i, k — случайные

$F \quad = \quad choose \quad M \quad best \quad of \quad F$

$p_f \quad = \quad choose \quad one \quad best \quad of \quad F$

8 Алгоритм построения модели глубокого обучения как суперпозиции локальных моделей

Вычислительный эксперимент рассматривает два подхода к построению модели: базовый и предложенный в работе. Базовый подход к объединению вышеописанных процедур построения модели подразумевает последовательное выполнение базовых процедур сегментации (6), группировки (7) и порождения признаков (8) и достижение оптимума локальных критериев качества для каждой из процедур. При этом описанные критерии построения процедур не связаны напрямую с целевым функционалом, что приводит к неоптимальности построенного решения.

Выше, в разделах 5, 6 и 7 описаны предлагаемые процедуры оптимизации соответствующих процедур. Принцип их объединения изложен в разделе 4. Последовательно выполняются следующие шаги:

1) инициализаци начальных точек для градиентных методов и начальных поколений для генетических алгоритмов,

2) шаги градиентного метода оптимизации для процедуры сегментации для каждого признака,

3) итерация генетической оптимизации для процедуры группировки для каждого признака,

4) итерация генетической оптимизации для процедуры порождения,

5) возврат ко 2 шагу до сходимости функционала качества.

Ниже приводится псевдокод объединения соответствующих процедур в едином алгоритме:

$for \quad iteration = 1,\dots,N:$

$\quad for \quad \chi_i, \quad i \in \mathscr{A}_l:$

$\quad\quad p_s^i = p_s^{i-1} - \lambda \bigtriangledown Q(p_s)$

$\quad\quad p_s^i = Correction(p_s^i)$

$\quad for \quad \chi_i, \quad i \in \mathscr{A}_c:$

$\quad\quad for \quad j = 1,\dots,N:$

$\quad\quad\quad A \quad append \quad crossing(a_\chi^i, a_\chi^k)$, где i,k — случайные

$\quad\quad\quad A \quad = \quad select \quad M \quad best \quad of \quad A$

$\quad\quad\quad p_g^i \quad = \quad select \quad one \quad best \quad of \quad A$

$F \quad append \quad crossing(a^i, a^k)$, где i,k — случайные

$F \quad = \quad select \quad M \quad best \quad of \quad F$

$p_f^i \quad = \quad select \quad one \quad best \quad of \quad F$

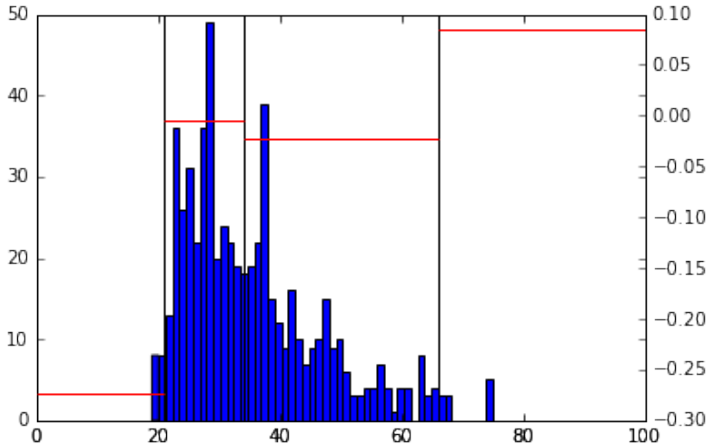

Рис. 7: Пример сегментации линейного признака и соответствующий вес в модели.

9 Вычислительный эксперимент

Построенная процедура оптимизации всех блоков в единой структуре модели сравнивается набором криетриев с другими моделями: нейронные сети, случайные леса и бустинг над деревьями. Результатом вычислительного эксперимента является сравнение качества моделей без подбора гиперпараметров. Модели сравниваются по AUC для оценки их качества, по структурной сложности моделей для оценки их устойчивости, интерпретируемости и сложности. Интерпретируемость итоговой модели оценивается для предложенного метода.

Для проведения вычислительного эксперимента и демонстрации интерпретируемости результатов построенной модели выбраны датасеты по кредитному скорингу: данные репозитория UCI German Credit Data [18]. Из множества признаков в выборке выбрано подмножество, содержащее: линейные признаки, категориальные признаки и бинарные признаки.

Результаты оптимизации процедуры сегментации для нескольких линейных признаков приведены на графиках 7–9. На этих графиках изображено распределение линейных признаков в виде гистограммы, а также предложенные параметры сегментации признака как узлы сегментации и веса соответствующих бинарных признаков принадлежности сегменту в модели для каждого из сегментов. Видно, что разбиение линейных признаков на группы интерпретируемо. Важно объединение хвостов распределения в одну категорию для ее интерпретируемости из-за малого числа объектов.

Из результатов моделирования процедуры сегментации можно сделать выводы о разделении заемщиков по социальным группам и склонности каждой из групп к кредитному риску. Например, из графика 8 видно, что молодые и пожилые люди с точки зрения банка относятся к более рискованным группам.

Для оценки сходимости оптимизационных процедур генетических алгоритмов был проведен ряд экспериментов: рассмотрены отдельно результаты моделирования процедуры группировки и процедуры порождения признаков. Ниже приведены примеры, демонстрирующие построенное решение:

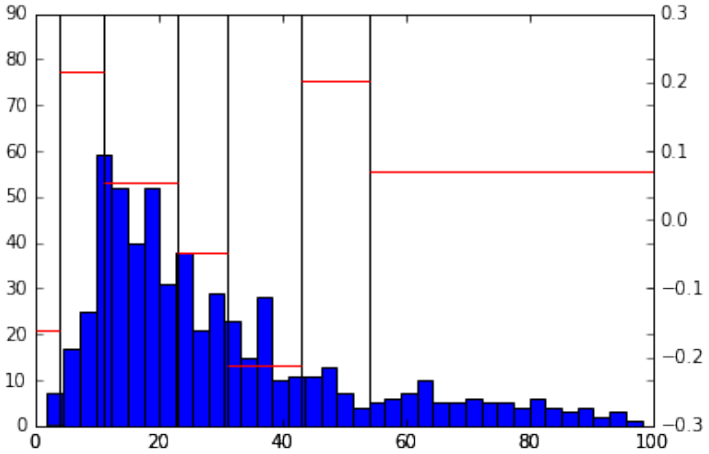

Рис. 8: Пример сегментации линейного признака и соответствующий вес в модели.

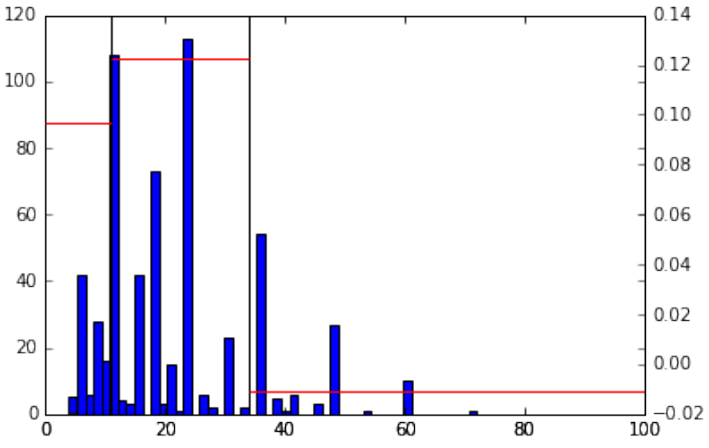

Рис. 9: Пример сегментации линейного признака и соответствующий вес в модели.

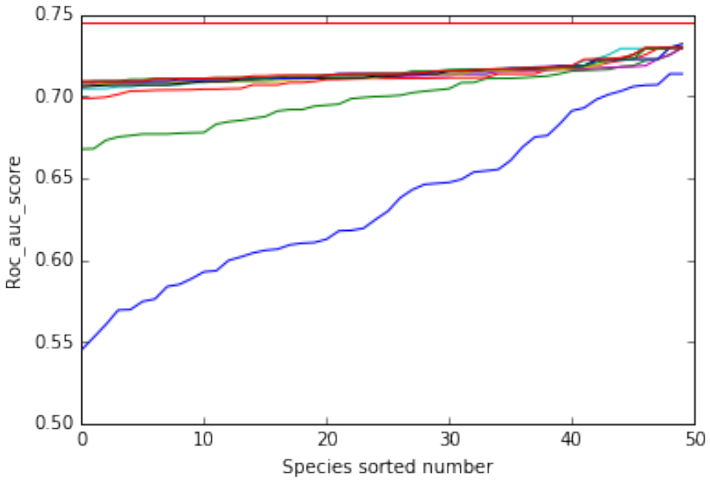

Рис. 10: Сходимость процедуры оптимизации группировки.

$$\chi = 1\ 2\ 3\ 4\ 5 \qquad 5 \text{ — число исходных категорий}$$
$$\downarrow\ \downarrow\ \downarrow\ \downarrow\ \downarrow$$
$$\chi_h = 1\ 2\ 2\ 3\ 3 \qquad 3 \text{ — число новых категорий, } \gamma_i \in h_{gi}$$
$$\chi = 0\ 1\ 2\ 3\ 4\ 5 \qquad 6 \text{ — число исходных категорий}$$
$$\downarrow\ \downarrow\ \downarrow\ \downarrow\ \downarrow\ \downarrow$$
$$\chi_h = 3\ 1\ 3\ 2\ 4\ 1 \qquad 4 \text{ — число новых категорий, } \gamma_i \in h_{gi}$$

Гиперпараметры модели подобраны автоматически согласно описанной теории. При генерации поколения присутствовали особи с различным числом категорий. Наилучшее качество доставляется оптимальным значением гиперпараметров. Объединение категорий приводит к повышению утойчивости моделей, их интерпретируемости и качества итогового решения. В предложенном наборе данных отсутствуют категориальные признаки с большим числом категорий, поэтому эффект от снижения их числа незначителен.

Ниже на графике кажда линия демонстрирует упорядоченное по качеству отдельной особи поколение. По оси абсцисс отложен упорядоченный по качеству номер особи в одном поколении. Сходимость максимума графика к стабильному числу указывает на сходимость и повышение качества итоговой лучшей особи, которая и выбирается как оптимальный параметр процедры группировки. Видно, что качество решения быстро сходится к оптимальному значению, а наилучшее задает максимальное качество.

Анализируя полученные разбиения исходных признаков видим, что разбиение, на котором достигает максимум функционала качества, является интерпретируемым. Например, признак кредитной истории, содержащий 5 категорий разбивается на три категории: ”Просрочил платеж в прошлом“ и ”Критический аккаунт“, ”Не брал“, ”Выплачивал“ и ”Выплачивал до настоящего момента“, что является интерпретируемым результатом: негативные примеры объединились в одну категорию, а позитивные и нейтральные - в другие две.

Аналогичный эксперимент проведен для процедуры порождения признаков. Каждая линия на

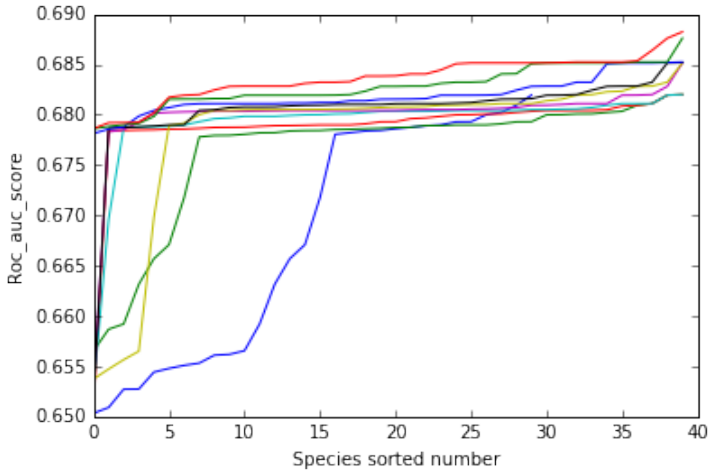

Рис. 11: Сходимость процедуры оптимизации порождения признаков.

Модель	Качество ($\mathscr{L}$) Roc-Auc	Качество ($\mathscr{T}$) Roc-Auc	Число параметров
Logistic Regression	0.751	0.704	6
End-to-end LR	0.782	0.730	25
XGBoost	0.912	0.729	5000
NN	0.834	0.720	3000

Таблица 1: Сравнение качества работы алгоритмов

графе показывает качества одного поколения. Видно, что лучшая особь каждого поколения имеет возрастающее значение функционала качества. Проанализируем построенные признаки, например: $\chi_1' = \frac{1}{\chi_1+\chi_2+\chi_3}$, $\chi_2' = \sqrt{\chi_2+\sqrt{\chi_1+\chi_3}}$. Не все из этих признаков являются интерпретируемыми, существует несогласованность операций и размерностей: складываются, например, кол-во месяцев на текущей работе с кол-вом взятых ранее кредитов.

Решение обладает высокой степенью интерпретируемости. В таблице 1 сравнивается качество и сложность моделей. Из результатов вычислительного эксперимента видно, что построенное решение применимо, его качество при малой сложности признакового пространства и пространства параметров и гиперпараметров остается на высоком уровне, уровне моделей со сложной и неинтерпретируемой структурой. Такой результат был достигнут засчет построенной процедуры порождения и изменения признакового пространства.

10 Заключение

В работе предложена единая процедура преобразования признакового пространства в задаче кредитного скоринга, объединяющая известные интерпретируемые процедуры обработки признакового

пространства. Показана работоспособность описанного подхода и продемонстрированы результаты моделирования на открытых данных. Продолжение работы предполагает добавление процедуры отбора признаков и усложнение процесса порождения признакового пространства. Также в дальнейших исследованиях планируется построение единой адаптивной онлайн-процедуры порождения мультимоделей кредитного скоринга над полученным признаковым пространством.

11 Список Литературы

[1] A. Kraus. Recent Methods from Statistics and Machine Learning for Credit Scoring // Dissertation an der Fakultat fur Mathematik, Informatik und Statistik der Ludwig–Maximilians–Universitat, Munchen. 2014.

[2] Y. Dong. A Case Based Reasoning System for Customer Credit Scoring: Comparative Study of Similarity Measures // Proceedings of the 51st Annual Meeting of the ISSS. 2007. Tokyo, Japan.

[3] С. В. Уланов. Оценка качества и сравнение скоринговых карт // Математические и инструментальные методы экономики. Экономические науки. 2009. 9(58).

[4] I. Oliveira, M. Chari and S. Haller. Rigorous Constrained Optimized Binning for Credit Scoring // SAS Global Forum. 2008.

[5] N. Siddiqi. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring // New Jersey: John Wiley Sons. 2006.

[6] USDOJ. Title VII—Equal Credit Opportunity Act // http://www.usdoj.gov/crt/housing/documents/ecoafulltext_5- 1- 06.htm. 2006.

[7] S. Scitovski and N. Sarlija. Cluster analysis in retail segmentation for credit scoring // CRORR 5. 2014. 235–245

[8] L. J. Sanchez-Barrios, G. Andreeva and J. Ansell. Time-to-profit scorecards for revolving credit // European Journal of Operational Research. 2016. Vol. 249, Iss. 2. Pp. 397–406.

[9] S. Maldonado, J. Pereza and C. Bravo. Cost-based feature selection for Support Vector Machines: An application in credit scoring // European Journal of Operational Research. 2017. Vol. 261, Iss. 2. Pp. 656–665.

[10] C. Luo, D. Wu and D. Wu. A deep learning approach for credit scoring using credit default swaps // Engineering Applications of Artificial Intelligence. 2017. Vol. 65. Pp. 465–470.

[11] X. Yufei, L. Chuanzhe, L. Y. Ying and L. Nana. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring // Expert Systems with Applications. 2017. Vol. 78. Pp. 225–241.

[12] X. Chen, C. Zhou, X. Wang and Y. Li. The Credit Scoring Model Based on Logistic-BP-AdaBoost Algorithm and its Application in P2P Credit Platform //Proceedings of the Fourth International Forum on Decision Sciences. 2017. Pp. 119–130

[13] G. Zeng. Invariant properties of logistic regression model in credit scoring under monotonic transformations // Communications in Statistics — Theory and Methods. 2017. Vol. 46. Iss. 17. Pp. 8791–8807.

[14] Y. Wang and J. L. Priestley. Binary Classification on Past Due of Service Accounts using Logistic Regression and Decision Tree // Grey Literature from PhD Candidates. 2017. Vol. 4. http://digitalcommons.kennesaw.edu/dataphdgreylit/4

[15] S. Walusala W, R. Rimiru and C. Otieno. A hybrid machine learning approach for cregit scoring using PCA and logistic regression // International Journal of Computer. 2017. Vol. 27. Iss. 1.

[16] A. Mathew. Cregit scoring using logistic regression // San Jose State University. 2017. Master's Projects. 532. http://scholarworks.sjsu.edu/etd_projects/532

[17] R. J. Connor. Grouping for Testing Trends in Categorical Data // Journal of the American Statistical Association. 1972. Vol. 67. Iss. 339.

[18] Data. Statlog (German Credit Data) Data set. https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)