Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Кузнецов Михаил Константинович

Автоматическая генерация признаков на табличных данных

Automatic feature generation for tabular data

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:
д.ф.-м.н., профессор
А. Г. Дьяконов

Москва, 2022

Содержание

§1. Введение

Машинное обучение (machine learning) применяется в большом количестве областей, например, в здравоохранении [1], финансах [2], мониторинге окружающей среды [3], мягкой робототехнике [4]. Чаще всего процесс решения конкретной задачи выглядит следующим образом: сбор данных, их предобработка, генерирование новых признаков, выбор критерия качества, выбор модели и ее обучение. В зависимости от того, насколько простая будет связь между признаками и целевой меткой, настолько модель будет быстрее обучаться и иметь меньшую сложность.

Из-за экспоненциального роста числа всевозможных трансформаций над признаками задача построения полезных признаков трудозатратна. Появляется естественное желание автоматизировать этот процесс. Одна из ключевых подзадач в области автоматического машинного обучения (AutoML) [5] — умная генерация признаков.

В данной работе описываются методы автоматической генерации признаков на табличных данных, проводятся эксперименты как на реальных, так и на искусственных данных.

§2. Обзор литературы

Основное деление методов автоматической генерации признаков:

- генерация-отбор или итеративный (expand reducion):
  - все сразу генерируем признаки [6], [7], [8],
  - только перспективные [9].
- генетический [10], [11],
- с помощью дерева [12], [13], [14], [15],
- с помощью нейросетей [16], [17], [18].

Деление трансформаций, используемых для получения новых признаков:

1. Унарные:
   - применение нелинейных функций: exp(x), log(1 + x).
2. Бинарные:
   - +, -, /, *,
   - агрегация (groupby + func).
3. Многоуровневые:
   - кодирование категориальных признаков (целевое кодирование (TE), количественное кодирование (count encoding)),

- методы снижения размерности:
  - автокодировщиком (AutoEncoder),
  - методом главных компонент (PCA) [19],
  - путем максимизации непараметрической взаимной информации (MMI) [20].
- knn признаки в задаче без учителя (расстояния между наблюдениями, признаки ближайших наблюдений),
- предсказание одних признаков по другим,
- показатель качества / функция потерь восстановленных данных с помощью других моделей (плотность, mse),
- кластеризация,
- взаимодействия категориальных признаков:
  - декартово произведение (кросс-признаки, crossover),
  - M-of-N — истина, когда хотя бы M из N условий верны,
  - ДНФ, оставляем важные исходя из коэффициентов Фурье [21].

Большинство методов, основанных на генерации-отборе, имеют встроенный отборщик признаков, в качестве которого выступают:

- взаимная информация,
- корреляция и корреляция расстояния (distance correlation) [22],
- веса линейной модели,
- оценка важности мета-моделью.

Например, в ExploreKit [6] авторы обучают мета-модель на большом объеме данных и предсказывают вероятность того, что данная трансформация признака даст прирост в качестве. В данном методе итоговое качество признака меряется как разница качества на исходном датасете с признаком и без него.

Генетический метод Few [10] в отличии от метода главных компонент имеет сложность, зависящую линейно от количества объектов в выборке. Это позволяет применять его на больших данных. FCDRGP [23] полезен для бустинга, так как он подбирает признаки исходя из показателя качества, основанного на разделении классов, так называемого fitness function. Суть в том, что если мы засэмлируем часть объектов из класса “c” и посмотрим какие объекты лежат недалеко от центра выборки, то в «хорошем» случае мы увидим объекты того же класса “c”, то есть энтропия должная быть низкая. Остаются признаки с относительно большим fitness function.

Преимущество SAFE [6] подхода перед итеративными заключается в сложности, зависящей от квадрата логарифма числа признаков, а не от квадрата числа признаков.

AutoCross [9], AutoFIS [24], AutoGroup [25], AutoINT [26], DNN2LR [18], FIVES [17], SIGN [27] предназначены для получения только кросс-признаков. Чаще всего в них

количественные признаки дискретизируется, то есть конвертируются в категориальные. Нейросетевые архитектуры популярны для задачи предсказания вероятности клика на событие, поэтому в качестве функции потерь используется бинарная кросс-энтропия. AutoFIS пытается найти коэффициенты при интеракциях второго порядка. Для правильного шкалирования и дополнительной устойчивости используется батч-нормализация скалярного произведения эмбеддингов. Итого, обучается совместно веса и коэффициенты с помощью GRDA оптимизатора [28], который способствует разряженности коэффициентов. После шага оптимизатора коэффициенты бинаризуются и обучение повторяется. В AutoGroup проблема выбора конкретной группы признаков решается с помощью Гумбель-Софтмакс трюка [29], в котором необходимо обучать вектор параметров — альфа. Внутри группы признаки складываются с коэффициентами, зависимыми от признака и порядка интеракции. Интеракции строятся по аналогии с FM [30], только обобщаются на большую степень. В качестве входа для последнего перцептрона (mlp) подается конкатенация получившихся представителей групп всех порядков. Обучаются поочередно альфа и параметры сети, включая коэффициенты признаков для каждого порядка, по всему датасету.

AutoINT использует трансформер [31] для извлечения попарных взаимодействий признаков. Эксперименты авторов AutoINT показали, что показатель качества auc roc (площадь под кривой ошибок) растет, как корень из количества слоев или размерности скрытого представления (hidden state). DNN2LR основан на подсчете нелинейной связи признаков по градиентам нейронной сети. После бинаризации по квантилю оценок нелинейности, авторы оставляют часто встречаемые комбинации признаков. Далее логистическая регрессия отбирает из оставшихся признаков нужное количество. Данный подход на порядок быстрее AutoCross.

Основная идея FIVES — поиск кросс-признаков с помощью нахождения матрицы смежности графа. Вершины графа — признаки (первого, второго и т.д. порядков). Ребро между вершинами графа — вхождение признаков в кросс-признак. При обучении использовалось температурное шкалирование, решающее проблему неуверенности в проведении ребра. Исходя из экспериментов, поставленными авторами FIVES, AutoFIS хуже FIVES статистически значимо. SIGN похож на FIVES, но можно отметить несколько различий. Первое, обучение классификатора на каждой итерации построения интеракций большего порядка, а не только на последней итерации. Второе, отдельный перцептрон, получая на вход две вершины, выдает вероятность проведения ребра. Третье, добавление l_0 регуляризации на вероятности появления ребра, которая аппроксимируется через hard concrete распределение [32] для возможности дифференцирования.

В [33] каждая вершина, за исключением начальной, представляет трансформированный датасет. Вводятся следующие определения:

- состояние (state) — трансформированный граф и оставшееся ограниченное количество возможных трансформаций графа,
- действие (action) — трансформация вершины, т.е. применение операции ко всевозможным признакам или парам признаков в данном датасете,
- награда на i-ом шаге (reward) — разница максимальных показателей качества модели, обученной на графе на $(i+1)$-ом и i-ом шаге.

LAFEM [14] обобщает подход [33], переопределяя понятие состояние:

- состояние (state) — трансформированный граф или признак и оставшееся ограниченное количество возможных трансформаций графа.

Кроме того, исходных вершин может быть несколько. Для ускорение работы сделаны следующие изменения: применение трансформаций не ко всевозможным парам признаков, подсчет показателя качества не во всех вершинах. CAFEM [13] помимо широкоизвестных трансформаций использует расширенное квантильное представление признака (extended quantile sketch array). Вершиной в CAFEM может быть только признак. Метаобучение политики на большом объеме датасетов помогает быстрее находить хорошие трансформации признаков на новых данных.

§3. Постановка задачи

Пусть выборка обозначается как $D = (x_i, y_i)_{i=1}^n$, где $(x_i, y_i) = ((x_{i1}, x_{i2}, ..., x_{ip}), y_i)$. $X = (x_i)_{i=1}^n$, $Y = (y_i)_{i=1}^n$. Допустим мы обучаем модель $\mathbf{a}$. Пара $(D, \mathbf{a})$ формирует конкретную ситуацию. Обозначим за $\mathcal{S}$ набор из всевозможных таких пар.

Тогда задача в общем случае выглядит следующим образом: необходимо задать функцию $\phi : \mathcal{S} \to \mathbb{R}^{n \times d}$, которая возвращает трансформированные данные.

Естественно, интерес заключается в нахождении новых признаков, которые:

- улучшают показатель качества решения итоговой задачи,
- быстро считаются,
- полезны для моделей машинного обучения и задач, отличных от пары $(D, \mathbf{a})$.

В частном случа есть набор трансформаций $\mathrm{tr}_1, \mathrm{tr}_2, ..., \mathrm{tr}_\mathrm{N}$ над признаками. Последовательность трансформаций $\mathrm{tr}_{\mathrm{i}_1}, \mathrm{tr}_{\mathrm{i}_2}, ..., \mathrm{tr}_{\mathrm{i}_\mathrm{m}}$ над исходными признаками называется формулой. Задача состоит в нахождении оптимального с точки зрения показателя качества модели $\mathbf{a}$ конечного набора формул, размер которого ограничен.

§4. Методы решения

4.1. Autofeat

Авторы данной статьи [7] сначала применяют классический подход: генерация унарных трансформаций признаков и затем использование бинарных трансформаций. Например, в качестве унарной трансформации взяли sin(x), а в качестве бинарной “+”. Для отбора признаков применялись следующие методы:

1. Использование веса модели линейной регрессии в качестве важности признака: чем меньше модуль веса, тем меньше важность.
2. Корреляция с другими признаками должна быть меньше определенного порога.
3. Для поиска полезных признаков брались подмножества трансформированных признаков, в которые добавлялись зашумленные важные признаки с шага 1. Если вес какого-то признака оказался больше веса зашумленного, такой признак считался «хорошим».
4. Отбор на основе длины формулы. Чем меньше, тем лучше.

Отбор происходит итеративно: на каждом шаге генерируются и отсеиваются ненужные признаки.

4.2. Neural Feature Search (NFS)

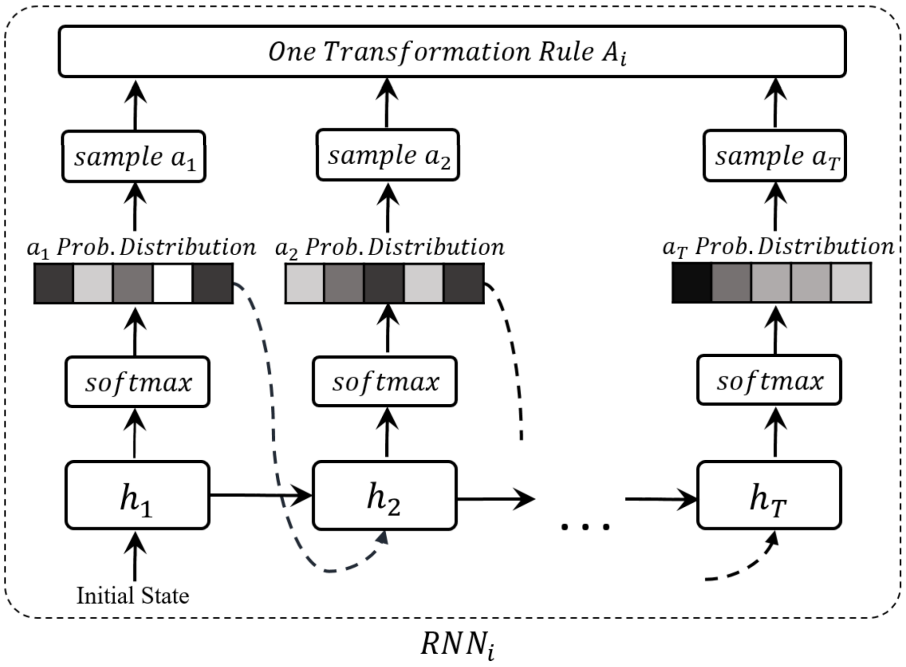

Рис. 1: Схема работы нейронной сети для генерации трансформаций в методе nfs [16]

Представим, что для каждого признака у нас есть модель, которая предсказывает какую трансформацию стоит сделать, то есть в конце модели будет стоять софтмакс

(softmax) преобразование, дающее нам вероятности полезности трансформаций. Трансформации большего порядка T можно получить если подавать на вход текущей модели выход предыдущей (см. рис. 1).

Policy gradient [34] алгоритм используется для обучения модели. В качестве состояния (state) выступает набор вероятностей трансформаций на каждом временном этапе. Чтобы получить награду (reward) от конкретной трансформации (action), считается разница показателей качества основной модели на датасетах с трансформированными признаками на шаге t и $t-1$:

$$R_t = Q_t - Q_{t-1}.$$

Дисконтированная кумулятивной награда считается, как

$$R_t^{(k)} = R_t + \gamma R_{t+1} + \ldots + \gamma^k R_{t+k}.$$

Пусть

$$R_t^{\lambda} = (1-\lambda)\sum_{k=1}^{p\times T} \lambda^{k-1} R_t^{(k)},$$

тогда оптимизируется функционал качества

$$L(\theta) = -\mathbb{E}_{P\left(a_{1:p\times T};\theta\right)}\left[\sum_{t=1}^{p\times T} R_t^{\lambda}\right],$$

где θ — параметры модели. Далее применяется REINFORCE [35] правило и Монте-Карло оценка [36] для получения аппроксимации градиента. Так как количество моделей зависит линейно от количества признаков данный подход не применим к задачам с большим количеством признаков.

4.3. Scalable Automatic Feature Engineering Framework (SAFE)

Один из распространенных подходов получения новых признаков – взятие их из модели машинного обучения. Так, авторы SAFE с помощью бустинга находят полезные комбинации признаков и используют их для создания новых.

В частности, считается разница информационных энтропий (information gain ratio) узла-родителя и узлов-детей для каждого пути в дереве. Если разница большая, то такую комбинацию признаков сохраняем и делаем над ними трансформации. Для отбора признаков используется information value (IV) после бинаризации по порогу:

$$\mathrm{IV} = \sum_i \left(\frac{n_p^i}{n_p} - \frac{n_n^i}{n_n}\right) \times \frac{n_p^i/n_p}{n_n^i/n_n},$$

где n_p, n_n — число положительных и отрицательных меток; n_p^i, n_n^i — число меток соответствующего класса, попавших в i-й бин. Если IV мало, такой признак отбрасывается. Слишком скоррелированные признаки также убираются.

4.4. AutoLearn

Предлагаемая авторами модель [37] в качестве отборщика признаков использует взаимную ифнормацию (Mutual Information) и корреляцию расстояния (distance correlation). Взаимная информация равна:

$$\mathrm{MI}(X;Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbb{P}_{X,Y}(x,y)\log_2\frac{\mathbb{P}_X(x)\mathbb{P}_Y(y)}{\mathbb{P}_{X,Y}(x,y)},$$

где $\mathbb{P}_{X,Y}(x,y)$ — вероятность встретить объект (x,y) в выборке, $\mathcal{Z}$ — множество значений переменной Z. Пусть

$$a_{i,j} = \|x_i - x_j\|_2, b_{i,j} = \|y_i - y_j\|_2, i,j = 1,2,\ldots,n,$$
$$A_{i,j} := a_{i,j} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}, \quad B_{i,j} := b_{i,j} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot},$$

где $\bar{a}_{i\cdot}$ — среднее значение i-ой строчки, $\bar{a}_{\cdot j}$ — среднее значение j-го столбца.

$$\mathrm{Cov}_{\mathrm{n}}^2(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}B_{i,j}$$

Тогда квадрат корреляции расстояния равен

$$\mathrm{Cor}_{\mathrm{n}}^2(X,Y) = \begin{cases} \frac{\mathrm{Cov}_{\mathrm{n}}^2(X,Y)}{\sqrt{\mathrm{Cov}_{\mathrm{n}}^2(X,X)\,\mathrm{Cov}_{\mathrm{n}}^2(Y,Y)}}, & \mathrm{Cov}_{\mathrm{n}}^2(X,X)\,\mathrm{Cov}_{\mathrm{n}}^2(Y,Y) > 0 \\ 0, & \mathrm{Cov}_{\mathrm{n}}^2(X,X)\,\mathrm{Cov}_{\mathrm{n}}^2(Y,Y) = 0 \end{cases}$$

Алгоритм получения новых признаков состоит из следующих четырех основных этапов:

1. Предварительная обработка: оставляются только признаки с большой взаимной информацией (MI).
2. Определение скоррелированных признаков с помощью distance correlation. Данная оценка важности признака позволяет измерять любую нелинейную зависимость между признаками.
3. Построение новых признаков: предскажем моделью один признак по-другому, а также добавим разницу между предсказаниями и истинной меткой, как признак. Если два исходных признака имеют большую distance correlation, то в качестве модели используется ядерная гребневая регрессия, иначе обычная гребневая регрессия.
4. Отбор признаков делается с помощью случайного алгоритма лассо регрессии (RandomizedLasso) и использования взаимной информации по аналогии с п. 1.

4.5. LightAutoML (LAMA)

В данном методе берутся взаимодействия признаков, исходя из степени их кардинальности. После чего делается кодирование категориальных переменных целевым признаком (TE).

4.6. Sure Independence Screening and Sparsifying Operator (SISSO)

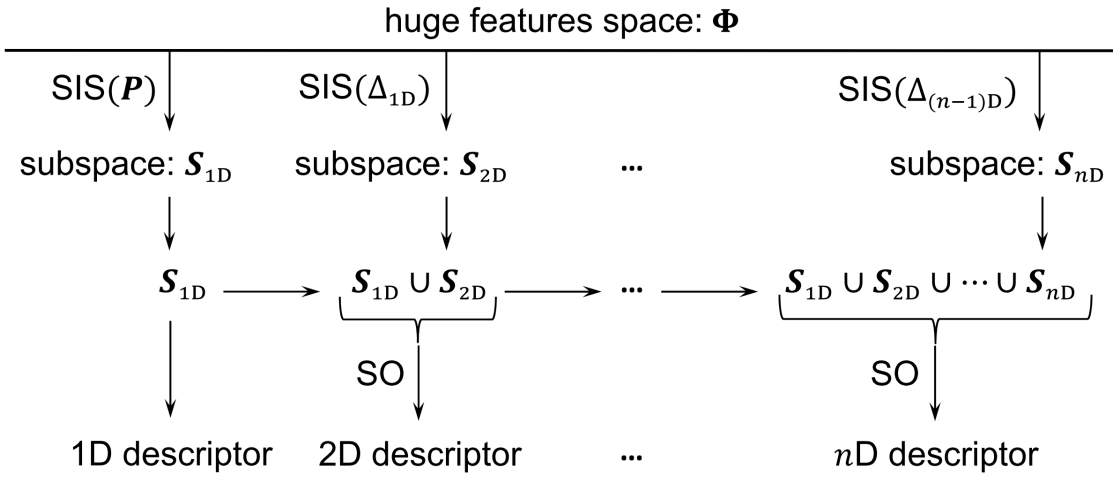

Рис. 2: Схема работы метода sisso [8]

Авторы данной статьи [8] изначально генерируют всевозможные трансформации признаков заданного порядка (Ф см. рис. 2). После чего стоит задача отбора признаков в нужном количестве. Для промежуточной стадии отбора используется корреляция (SIS) с целевым признаком (P). Алгоритм нахождения новых признаков выглядит так:

1. Для первой итерации выделяется группа «хороших» признаков (S_{1p}) и из них лучший (1d descriptor).
2. Подсчитывается ошибка гребневой регрессии ($\Delta_{i\,\mathrm{D}}$) при предсказании целевого признака с помощью признаков, выступающих в качестве кандидатов.
3. Берутся топ признаков высоко скоррелированных с ошибкой на шаге 2. Данные признаки сохраняются (S_{ip}).
4. Из признаков $S_{(i-1)\,\mathrm{D}} \cup S_{ip}$ с помощью l_0 или l_1 регуляризации отбираются i признаков.

Таким образом алгоритм постепенно находит подпространства, наиболее связанные с целевым признаком.

4.7. Feature Engineering Wrapper (FEW)

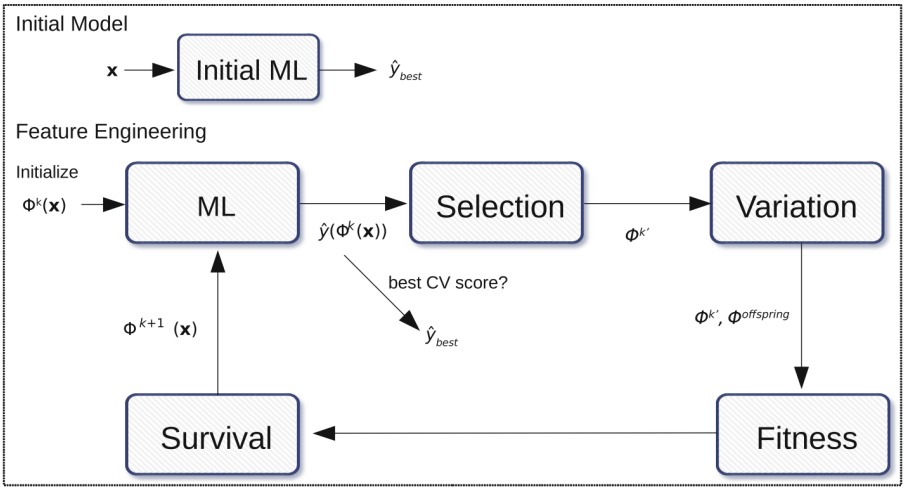

Рис. 3: Схема работы метода few [10]

Генетические алгоритмы применимы для получения новых признаков, если определить как происходит мутация и отбор поколений. Авторы данного метода использовали следующий подход (см. рис. 3):

1. Генерируются случайные группы трансформированных признаков (поколения).
2. Отбираются часть поколений по качеству решения задачи с помощью модели машинного обучения.
3. Текущие поколения подвергаются мутации:
   - кроссовер для дерева (tree crossover, см. рис. 4),
   - точечная мутация (point mutation, см. рис. 5).
4. Оценивается качество поколений.
5. Отбираются поколения через эпсилон выживание (*eps*-lexicase survival).

Эпсилон выживание заключается в следующей процедуре:

1. Инициализация поколений G и множества S случайных выборок из данных.
2. Берутся топ поколений находящихся в *delta* окрестности по качеству на i-й выборки, принадлежащей S, где $delta = \text{median}(|metric - \text{median}(metric)|)$, $metric =$ массив всех показателей качества поколений на i-й выборке.
3. Переход на 2. Цикл прерывается по достижению нужного числа поколений или отсутствию новых выборок данных.

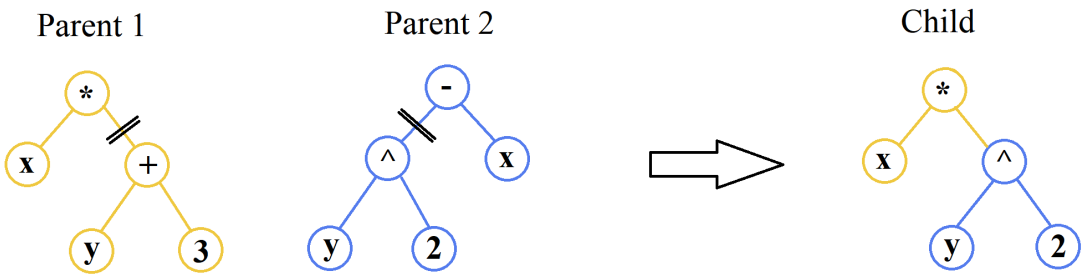

Рис. 4: Пример кроссовера для дерева.

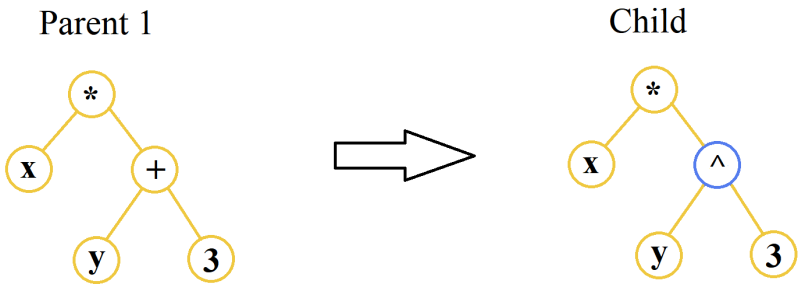

Рис. 5: Пример точечной мутации для дерева.

4.8. Iterative Feature Construction algorithm (TFC)

Данный алгоритм [38] заключается в итеративном построении новых трансформаций над уже существующими признаками и последующий их отбор с помощью теста хи-квадрат. В качестве трансформаций брались такие операции, как сложение, вычитание, умножение, деление. Исходные признаки могут быть полезны при комбинации с трансформированными, однако не гарантированно, что они пройдут многоэтапный отбор. Модифицированный метод выглядит следующим образом:

Algorithm 1 Модифицированный TFC

Вход:

$K \leftarrow$ количество итераций,

$(F, y) \leftarrow$ (признаки из датасета, целевая переменная),

$\mathrm{Tr} \leftarrow$ множество трансформаций,

$\chi^2 \leftarrow$ отбощик признаков на основе теста хи-квадрат,

$\mathrm{Tr}(F) \leftarrow$ применение трансформаций на всех парах признаков из F,

$\mathrm{MinMax}(F) \leftarrow$ применение мин-макс масштабирования F,

$\chi^2(F, y) \leftarrow$ отбор признаков из F.

Выход:

$F' \leftarrow$ новые признаки

Тело алгоритма:

$F' = F$

for $i = 1, ..., K$ **do**

$F' = \mathrm{Tr}(F')$

$F' = \mathrm{MinMax}(F')$

$F' = \chi^2(F', y)$

$F' = F' \cup F$

§5. Эксперименты

В качестве языка программирования для реализации методов автоматической генерации признаков и проведения экспериментов выбран python. Использовался docker для упрощения запуска экспериментов, так как не все методы исходно реализованы на python, например, safe написан на R. Часть методов, используемых или модифицированных нами, имеют определенные права, которые не позволяют их публиковать. Из-за этого код всех методов выложен на приватный github репозиторий [39]. Эксперименты вычислялись на машине с 80 ядрами и 1.5 терабайт оперативной памяти с ограничением в 4 дня на обучение одного метода.

5.1. Датасеты

Таблица 1: Датасеты для тестирования методов автоматической генерации признаков.

id	**name**	**abbr**	**nrow**	**ncol**	**ncat**	**nrel**	**source**
971	mfeat-fourier	MF	2000	76	0	76	autolearn
44	spambase	SB	4601	57	0	57	safe
979	waveform	WV	5000	40	0	40	autolearn
41146	sylvine	SL	5124	20	0	20	openml
1471	eeg-eye-state	ES	14980	14	0	14	safe
42477	credit-default	DF	30000	23	0	23	nfs
4135	amazon	AZ	32769	9	0	9	nfs
1461	bank-marketing	BM	45211	16	9	7	openml
41150	miniboone	MB	130064	50	0	50	openml
1169	airlines	AL	539383	7	3	4	openml

Часть датасетов была взята из открытого бенчмарка по автоматическому машинному обучению (OpenML AutoML) [40], другая часть из статьей. Если посмотреть на показатели датасетов (число строк, столбцов и тому подобное), то можно увидеть, что они сильно различаются, то есть данные репрезентативны. Для проведения эксперимента каждый датесет делится на обучающую (70%) и тестовую (30%) выборки. Рассматривается только задача бинарной классификации.

5.2. Предобработка данных

В данных содержатся количественные и категориальные признаки. Необходимо каким-то образом закодировать категориальные переменные, так как чаще всего они содержатся в виде текста. Для бустинга использовалось количественное кодирование, а для линейной регрессии one-hot. Перед обучением метода автоматической генерации признаков и линейной регрессии количественные признаки стандартизовались. Перед работой всех моделей и методов пропуски в данных заполнялись медианным значением.

5.3. Алгоритмы

Реализация авторов таких методов, как AutoLearn, SAFE — некорректная. В AutoLearn не итерируется один из счетчиков цикла, что приводит к созданию не всех трансформаций признаков. Кроме того, при предсказании значений признака на тестовых данных не используются данные из обучающей выборки. В SAFE обучение бустинга на валидационных данных делает модель переобученной.

Во всех экспериментах участвуют следующие модели:

- **бустинг (LightGBM)** с параметрами по умолчанию, за исключением параметров: bagging_freq=1, metric=auc roc, num_boost_round=10000, early_stopping_rounds=10
- **линейная модель (Logistic Regression)** с параметрами по умолчанию, за исключением параметра max_iter=1000.

Параметры выбраны так, чтобы модели не получились слишком простые. Обучение моделей происходило по кросс-валидации с пятью фолдами. Параметры оптимизировались с помощью optuna [41] на одном фолде с ограничениями в 100 попыток и 30 минут для максимизации показателя качества auc roc по следующей сетке гиперпараметров:

Таблица 2: Оптимизируемые гиперпараметры бустинга, нижняя и верхняя границы.

name	**low value**	**high value**
num_leaves	2	256
feature_fraction	0.4	1.0
bagging_fraction	0.4	1.0

Таблица 3: Оптимизируемый гиперпараметр логистической регрессии по логарифмической шкале, нижняя и верхняя границы.

name	**low value**	**high value**
C	1e-5	20.0

Подбор гиперпараметров (см. таблицы 2, 3) делает модели менее переобученными.

Критерий качества auc roc соответствует общепринятому показателю качества для задачи бинарной классификации на открытом бенчмарке по автоматическому машинному обучению [40]. В статьях [16], [42], [37] по автоматической генерации признаков еще используют сбалансированную точность, f1. Результаты экспериментов без оптимизации гиперпараметров для указанных показателей качества можно найти в приложении A. Далее auc roc обозначается как auc.

В качестве методов автоматической генерации признаков выступают:

- **AutoFeat**,
- **NFS**,
- **SAFE**,
- **AutoLearn**,
- **FEW**,
- **LAMA**,
- **TFC**.

Во всех методах использовались параметры из соответствующих статей.

5.4. Результаты

5.4.1 Ранги

	MF	SB	WV	SL	ES	DF	AZ	BM	MB	AL	rank
nfs	99.99	96.64	94.15	96.16	–	73.65	53.70	**91.04**	–	–	5.42
base	**100**	96.53	92.65	96.17	67.50	71.43	53.73	90.90	95.45	68.82	5.20
tfc	**100**	96.66	95.09	96.15	67.76	71.47	54.30	90.62	–	–	4.75
few	**100**	96.53	92.93	96.58	66.27	75.00	64.57	90.91	95.41	–	4.27
autolearn	**100**	**97.43**	95.12	96.08	67.50	71.43	53.73	90.90	95.45	68.82	4.25
lama	**100**	96.74	92.66	96.18	**75.47**	71.41	**85.83**	90.18	95.57	**71.44**	4.00
safe	**100**	96.70	94.92	97.29	60.22	74.50	60.45	90.70	96.34	68.40	4.00
autofeat	**100**	97.24	**95.28**	**97.44**	60.41	**75.70**	64.92	90.90	**97.47**	–	2.44

Таблица 4: Auc для модели linear на тестовых данных в зависимости от датасета и метода автоматической генерации признаков. Результаты представлены после подбора гиперпараметров моделей.

Линейная регрессия явно лучше работает на трансформированных признаках (см. таблицу 4). Специальный отбор в AutoFeat позволяет существенно повысить качество относительно других итерационных методов. NFS оказался на последнем месте, так как награда считалась, основываясь на показатель качества случайного леса. Наилучшим с точки зрения ранга оказался AutoFeat.

	MF	SB	WV	SL	ES	DF	AZ	BM	MB	AL	rank
safe	99.91	98.50	95.67	**99.16**	98.84	76.65	81.83	92.55	98.50	71.60	6.00
tfc	99.98	98.46	95.14	98.07	98.94	77.26	84.18	92.79	98.55	71.89	5.25
few	99.96	**98.85**	95.46	98.26	98.91	77.22	83.54	93.38	98.53	–	4.72
base	99.95	**98.85**	95.46	98.26	98.91	77.22	85.76	93.24	98.53	**72.22**	4.45
autolearn	98.31	98.71	95.09	98.59	99.04	77.47	85.76	93.24	98.53	**72.22**	4.34
autofeat	**99.99**	98.72	95.66	98.92	98.62	77.58	85.27	93.27	–	71.82	3.77
nfs	99.96	98.70	95.49	98.45	**99.13**	77.48	85.96	**93.60**	–	72.16	3.27
lama	99.98	98.80	**95.68**	98.32	98.91	**77.64**	**87.05**	92.48	**98.58**	72.19	3.05

Таблица 5: Auc для модели lgbm на тестовых данных в зависимости от датасета и метода автоматической генерации признаков. Результаты представлены после подбора гиперпараметров моделей.

В большинстве случаев среднее кодирование (LAMA) дает бустингу существенно больший прирост в качестве, нежели комбинации признаков, полученные после бинарных трансформаций (см. таблицу 5). Однако среднее кодирование может повысить шанс переобучиться, что не всегда дает уверенность в лучшем результате. Наилучшим с точки зрения ранга оказался метод LAMA.

5.4.2 Улучшение показателя качества

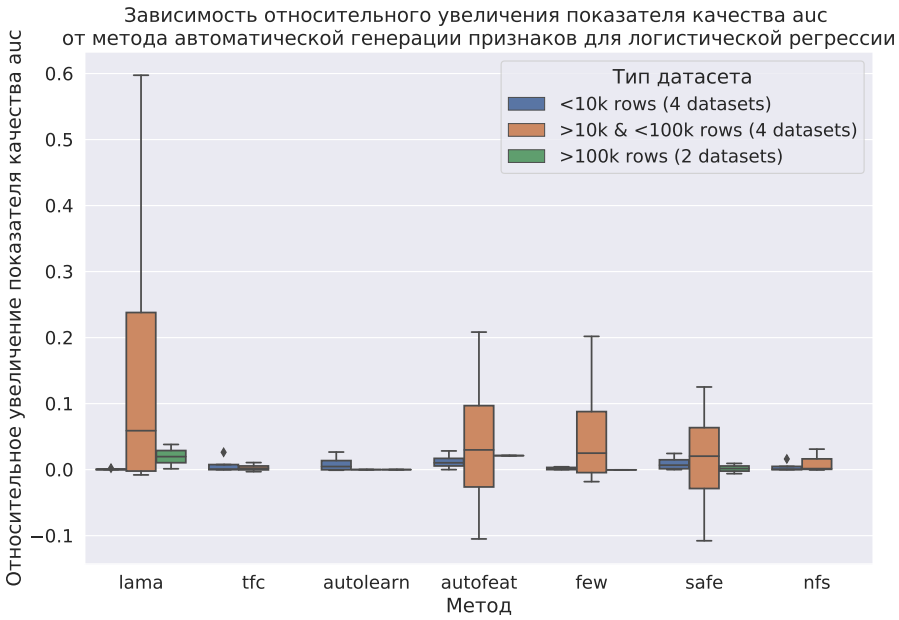

Рис. 6: Относительное улучшение показателя качества auc на 10 датасетах для логистической регрессии.

Улучшение показателя качества для линейной регрессии на значительный процент достигается на датасетах с размером строк, большим чем 10000 и меньшим чем 100000 (см. рис. 6). TFC и AutoLearn сильно хуже себя показали относительно других методов.

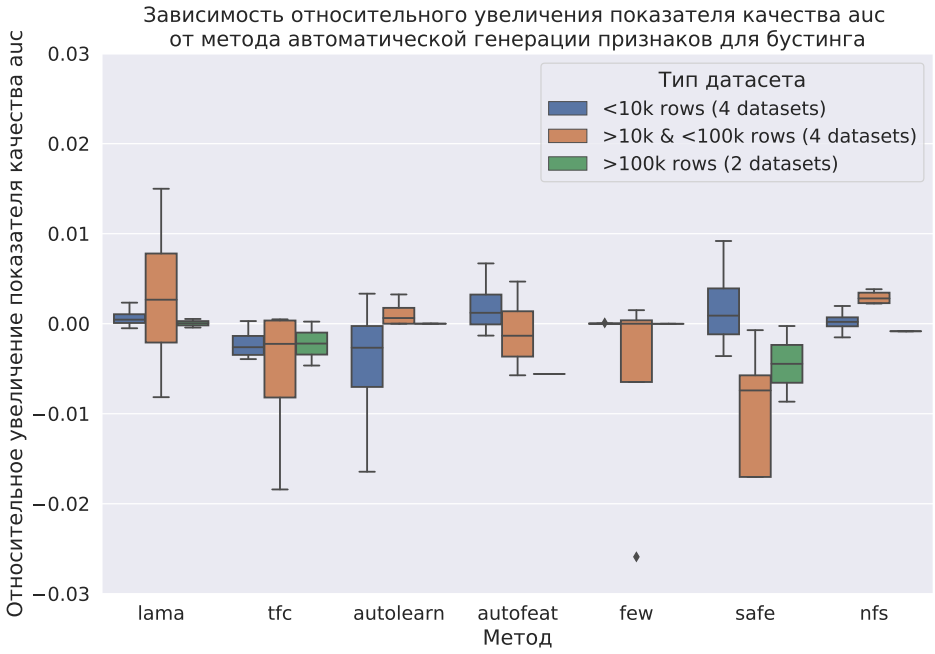

Рис. 7: Относительное улучшение показателя качества auc на 10 датасетах для бустинга.

Для бустинга сильного улучшения показателя качества не удалось достичь (см. рис. 7). SAFE плохо работает на датасетах с количеством строк, большим чем 10000.

5.4.3 Другие показатели

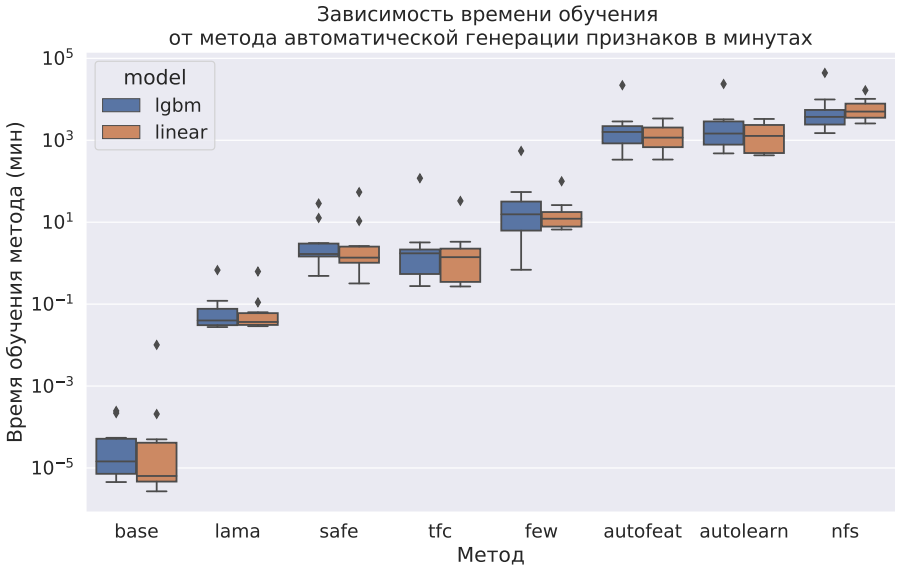

Рис. 8: Среднее время работы методов автоматической генерации признаков на 10 датасетах.

Дольше всего работает NFS, так как необходимо большое количество оценок показателя качества для подсчета награды (см. рис. 8). TFC на два порядка быстрее autolearn из-за отсутсвия необходимости построения регрессионных моделей.

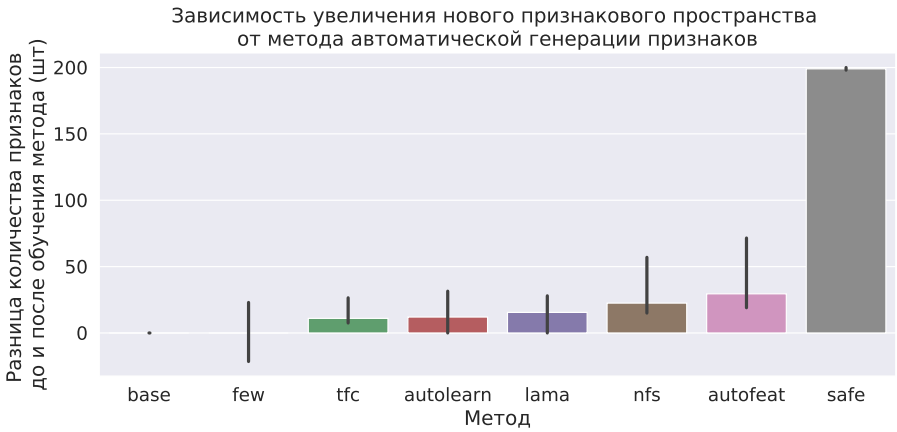

Рис. 9: Среднее увеличение количества признаков после применения метода автоматической генерации признаков на 10 датасетах. Высота столбца считается, как медиана. Отрезок возле вершины столбца — 95% доверительный интервал.

SAFE имеет самое большое число сгенерированных признаков и самую малую дисперсию (см. рис. 9). FEW в отличии от autolearn не конкатенирует исходной датасет к сгенерированным признакам, поэтому количество трансформированных признаков может уменьшаться.

§6. Заключение

В работе была рассмотрена задача автоматической генерации признаков на табличных данных. Были проведены эксперименты с двумя моделями машинного обучения и 7 методами генерации признаков. Основное улучшение качества удается добиться для логистической регрессии с помощью AutoFeat. На защиту выносится:

1. Реализация таких методов автоматической генерации признаков, как AutoLearn, TFC.
2. Доработка NFS, SAFE.
3. Добавление интерфейса на языке python ко всем использованным методам, а также возможности запуска эксперимента в docker окружении.
4. Проведение экспериментов на искусственных и реальных данных.

§7. Список литературы

1. *Waring J.*, *Lindvall C.*, *Umeton R.* Automated machine learning: Review of the state-of-the-art and opportunities for healthcare // Artificial Intelligence in Medicine. — 2020. — Vol. 104.
2. *Tiwari S.*, *Ramampiaro H.*, *Langseth H.* Machine Learning in Financial Market Surveillance: A Survey // IEEE. — 2021.
3. *Ghannam R. B.*, *Techtmann S. M.* Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring // Computational and Structural Biotechnology Journal. — 2021. — Vol. 19. — P. 1092–1107.
4. *Kim D.* Review of machine learning methods in soft robotics // PLoS One. — 2021. — Vol. 16, no. 2.
5. *He X.*, *Zhao K.*, *Chum X.* AutoML: A survey of the state-of-the-art // Knowledge-Based Systems. — 2021. — Vol. 212. — P. 106622.
6. *Katz G.*, *Shin E.*, *Song D.* Explorekit: Automatic feature generation and selection // IEEE 16th International Conference on Data Mining (ICDM). — 2016. — P. 979–984.
7. *Horn F.*, *Pack R.*, *Rieger M.* The autofeat python library for automated feature engineering and selection // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. — 2019. — P. 111–120.
8. *Ouyang R.* SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates // Physical Review Materials. — 2018. — Vol. 2, no. 8. — P. 083802.
9. *Luo Y.* Autocross: Automatic feature crossing for tabular data in real-world applications // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining. — 2019. — P. 1936–1945.
10. *La C.*, *Moore J.* General Feature Engineering Wrapper for Machine Learning Using ϵ-Lexicase Survival // European Conference on Genetic Programming. — 2017. — P. 80–95.
11. *Olson R.*, *Moore J.* TPOT: A tree-based pipeline optimization tool for automating machine learning // Workshop on automatic machine learning. – PMLR. — 2016. — P. 66–74.
12. *Shi Q.* Scalable automatic feature engineering framework for industrial tasks // IEEE 36th International Conference on Data Engineering (ICDE). — 2020. — P. 1645–1656.

13. *Zhang J.*, *Hao J.*, *Fogelman-Soulié F.* Cross-data Automatic Feature Engineering via Meta-learning and Reinforcement Learning // Pacific-Asia Conference on Knowledge Discovery and Data Mining. — 2020. — P. 818–829.

14. *Zhang J.* Automatic feature engineering by deep reinforcement learning // Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. — 2019. — P. 2312–2314.

15. *Khurana U.* Cognito: Automated feature engineering for supervised learning // IEEE 16th International Conference on Data Mining Workshops (ICDMW). – IEEE. — 2016. — P. 1304–1307.

16. *Chen X.* Neural feature search: A neural architecture for automated feature engineering // IEEE International Conference on Data Mining (ICDM). – IEEE. — 2019. — P. 71–80.

17. *Xie Y.* Fives: Feature interaction via edge search for large-scale tabular data // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining. — 2021. — P. 3795–3805.

18. *Liu Z.* Dnn2lr: Interpretation-inspired feature crossing for real-world tabular data // arXiv preprint arXiv:2008.09775. — 2020.

19. *Abdi H.*, *Williams L.* Principal component analysis // Wiley interdisciplinary reviews: computational statistics. — 2010. — Vol. 2, no. 4. — P. 433–459.

20. *Torkkola K.* Feature extraction by non-parametric mutual information maximization // Journal of machine learning research. — 2003. — Vol. 3. — P. 1415–1438.

21. *Duan J.* Automated generation and selection of interpretable features for enterprise security // IEEE International Conference on Big Data (Big Data). – IEEE. — 2018. — P. 1258–1265.

22. *Székely G.*, *Rizzo M.*, *Bakirov N.* Measuring and testing dependence by correlation of distances // The annals of statistics. — 2007. — Vol. 35, no. 6. — P. 2769–2794.

23. *Neshatian K.*, *Zhang M.*, *Johnston M.* Feature construction and dimension reduction using genetic programming // Australasian Joint Conference on Artificial Intelligence. – Springer. — 2007. — P. 160–170.

24. *Liu B.* Autofis: Automatic feature interaction selection in factorization models for click-through rate prediction // Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining. — 2020. — P. 2636–2645.

25. *Liu B.* AutoGroup: Automatic feature grouping for modelling explicit high-order feature interactions in CTR prediction // Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. — 2020. — P. 199–208.

26. *Liu B.* Autoint: Automatic feature interaction learning via self-attentive neural networks // Proceedings of the 28th ACM International Conference on Information and Knowledge Management. — 2019. — P. 1161–1170.

27. *Su Y.* Detecting beneficial feature interactions for recommender systems // dings of the 34th AAAI Conference on Artificial Intelligence (AAAI). — 2021.

28. *Chao S. K.*, *Cheng G.* A generalization of regularized dual averaging and its dynamics // arXiv preprint arXiv:1909.10072. — 2019.

29. *Jang E.*, *Gu S.*, *Poole B.* Categorical reparameterization with gumbel-softmax // arXiv preprint arXiv:1611.01144. — 2016.

30. *Rendle S.* Factorization machines // IEEE International conference on data mining. – IEEE. — 2010. — P. 995–1000.

31. *Vaswani A.* Attention is all you need // Advances in neural information processing systems. — 2017. — Vol. 30.

32. *Maddison C. J.*, *Mnih A.*, *Teh Y. W.* The concrete distribution: A continuous relaxation of discrete random variables // arXiv preprint arXiv:1611.00712. — 2016.

33. *Khurana U.*, *Samulowitz H.*, *Turaga D.* Feature engineering for predictive modeling using reinforcement learning // Proceedings of the AAAI Conference on Artificial Intelligence. — 2018. — Vol. 32, no. 1.

34. Policy gra- dient methods for reinforcement learning with function approximation / R. Sutton [et al.] // in Proceedings of NIPS. — 1999. — P. 1057–1063.

35. *Williams R. J.* Simple statistical gradient-following algorithms for connectionist reinforcement learning // Machine learning. — 1992. — Vol. 8, no. 3. — P. 229–256.

36. *Robert P. C.*, *Casella G.* Monte Carlo Statistical Methods // Springer Texts in Statistics. — 2004.

37. *Kaul A.*, *Maheshwary S.*, *Pudi V.* Autolearn—Automated feature generation and selection // IEEE International Conference on data mining (ICDM). — 2017. — P. 217–226.

38. *Piramuthu S.*, *Sikora R.* Iterative feature construction for improving inductive learning algorithms // Expert Systems with Applications. — 2009. — Vol. 36, no. 2. — P. 3401–3406.

39. *Kuznetsov M.* Accompanying repository: Automatic feature generation for tabular data // URL: https://github.com/MikhailKuz/afg. — 2022.

40. *Gijsbers P.* An open source AutoML benchmark // arXiv preprint arXiv:1907.00909. — 2019.

41. *Akiba T.* Optuna: A next-generation hyperparameter optimization framework // Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery data mining. — 2019. — P. 2623–2631.

42. *Dor O.*, *Reich Y.* Strengthening learning algorithms by feature discovery // Information Sciences. — 2012. — Vol. 189. — P. 176–190.

§A. Показатели качества

A.1. Без оптимизации гиперпараметров

В данном приложении приводятся такие показатели качества, как auc, f1, сбалансированная точность для всех датасетов и моделей.

	MF	**SB**	**WV**	**SL**	**ES**	**DF**	**AZ**	**BM**	**MB**	**AL**	**rank**
tfc	99.86	98.75	95.15	98.04	98.32	77.53	82.88	93.04	98.41	71.56	5.65
base	99.07	98.80	95.29	98.18	98.28	77.39	84.30	93.17	98.42	71.79	4.59
safe	**100**	98.58	**95.85**	**99.08**	98.57	77.25	80.78	92.59	**98.46**	71.34	4.50
lama	99.07	98.75	95.29	98.18	98.43	77.39	**87.13**	92.44	98.42	**72.32**	4.40
autofeat	99.98	98.73	95.25	98.91	97.95	77.52	85.58	93.23	–	71.66	4.11
few	99.89	98.80	95.29	98.18	98.28	77.39	85.36	93.27	98.42	–	4.11
autolearn	99.97	98.67	95.14	98.53	98.75	**77.81**	84.30	93.17	98.42	71.79	4.00
nfs	98.97	**98.86**	95.29	98.89	**98.91**	77.44	83.08	**93.62**	–	71.67	3.50

Таблица 6: Auc для модели lgbm на тестовых данных в зависимости от датасета и метода автоматической генерации признаков.

	MF	**SB**	**WV**	**SL**	**ES**	**DF**	**AZ**	**BM**	**MB**	**AL**	**rank**
tfc	**100**	96.59	95.08	96.15	62.98	71.47	54.30	90.54	–	–	5.43
base	**100**	96.68	92.75	96.17	62.99	71.43	53.73	90.81	93.63	68.82	5.00
nfs	**100**	96.56	94.15	96.16	–	73.65	53.86	**91.04**	–	–	4.92
autolearn	**100**	**97.71**	95.10	96.05	62.99	71.43	53.73	90.81	93.63	68.82	4.50
safe	**100**	96.66	94.22	96.89	54.66	74.52	62.31	90.20	96.23	68.40	4.25
lama	**100**	96.98	92.74	96.18	**74.70**	71.41	**85.53**	90.07	93.90	**71.44**	4.05
few	**100**	96.62	93.25	96.58	64.65	75.00	64.52	90.81	93.63	–	3.88
autofeat	**100**	96.94	**95.28**	**97.57**	63.48	**75.72**	64.94	90.81	**97.60**	–	2.22

Таблица 7: Auc для модели linear на тестовых данных в зависимости от датасета и метода автоматической генерации признаков.

	MF	SB	WV	SL	ES	DF	AZ	BM	MB	AL	rank
tfc	**96.61**	93.96	81.74	94.36	92.18	**54.23**	97.13	62.32	90.07	65.10	5.00
autofeat	0.00	94.58	82.63	95.73	91.75	53.44	97.30	61.99	–	65.14	4.50
safe	0.00	94.51	**83.11**	**95.97**	93.18	53.37	97.28	59.83	**90.23**	65.00	4.50
nfs	0.00	94.50	80.98	95.65	**94.18**	53.68	97.11	**63.07**	–	65.14	4.50
few	89.90	94.59	82.34	94.26	92.57	53.40	97.35	62.16	90.09	–	4.38
base	89.90	94.59	82.34	94.26	92.57	53.40	97.15	62.50	90.14	65.21	4.15
lama	89.90	**94.62**	82.34	94.26	92.72	53.40	**97.37**	60.79	89.94	**65.47**	4.05
autolearn	89.90	94.16	81.28	94.37	93.97	53.53	97.15	62.50	90.14	65.21	3.95

Таблица 8: F1 для модели lgbm на тестовых данных в зависимости от датасета и метода автоматической генерации признаков.

	MF	SB	WV	SL	ES	DF	AZ	BM	MB	AL	rank
nfs	**100**	90.20	80.27	91.76	–	51.56	97.01	58.08	–	–	5.71
safe	**100**	90.43	81.70	92.20	49.00	50.92	96.92	56.02	84.69	63.79	5.45
base	**100**	90.89	81.39	91.71	59.56	51.07	97.01	**58.17**	78.46	63.93	4.65
tfc	**100**	90.80	**83.88**	91.98	59.65	51.26	97.01	57.85	–	–	4.25
few	**100**	90.62	81.10	**92.87**	59.69	51.65	97.01	**58.17**	78.45	–	4.05
lama	**100**	91.05	81.20	91.85	**67.90**	50.82	**97.18**	56.56	79.99	**65.07**	4.05
autolearn	**100**	**92.19**	82.59	92.14	59.56	51.07	97.01	**58.17**	78.46	63.93	3.75
autofeat	**100**	91.59	83.77	92.77	61.19	**52.59**	97.01	**58.17**	**88.07**	–	2.38

Таблица 9: F1 для модели linear на тестовых данных в зависимости от датасета и метода автоматической генерации признаков.

	MF	SB	WV	SL	ES	DF	AZ	BM	MB	AL	rank
tfc	**96.85**	94.89	86.25	94.21	92.89	**71.26**	75.74	86.69	94.12	65.43	5.50
autofeat	50.00	95.54	86.71	95.64	92.52	70.55	78.39	87.13	–	65.53	4.55
base	95.00	**95.57**	86.66	94.08	93.27	70.56	76.87	86.76	94.20	65.66	4.34
autolearn	90.83	95.17	87.28	94.34	94.49	70.54	76.87	86.76	94.20	65.66	4.30
safe	50.00	95.54	**87.48**	**95.90**	93.81	70.58	74.13	85.71	**94.38**	65.20	4.30
lama	95.00	95.45	86.66	94.08	93.38	70.56	**80.01**	85.86	94.21	**66.03**	4.25
few	95.83	**95.57**	86.66	94.08	93.27	70.56	78.48	86.90	94.19	–	4.11
nfs	50.00	95.54	86.67	95.57	**94.72**	70.74	76.32	**87.48**	–	65.55	3.55

Таблица 10: Сбалансированная точность на тестовых данных для модели lgbm в зависимости от датасета и метода автоматической генерации признаков.

	MF	SB	WV	SL	ES	DF	AZ	BM	MB	AL	rank
safe	**100**	92.01	86.11	92.00	52.52	69.12	49.98	82.93	90.50	62.81	5.35
few	**100**	92.26	87.22	92.52	59.32	69.63	49.99	84.16	86.54	–	4.94
tfc	**100**	91.98	87.48	91.61	59.84	68.23	50.00	**84.20**	–	–	4.75
base	**100**	92.47	86.81	91.67	59.69	68.32	50.00	84.18	86.64	63.14	4.50
nfs	**100**	92.43	87.30	91.61	–	68.93	50.00	84.18	–	–	4.50
lama	**100**	92.61	86.40	91.35	**68.37**	68.55	**77.56**	82.46	87.12	**65.33**	4.15
autolearn	**100**	**93.68**	**89.02**	91.87	59.69	68.32	50.00	84.18	86.64	63.14	3.60
autofeat	**100**	93.01	87.74	**92.58**	60.68	**70.17**	50.00	84.18	**92.88**	–	2.33

Таблица 11: Сбалансированная точность на тестовых данных для модели linear в зависимости от датасета и метода автоматической генерации признаков.