Тема: "Исследование алгоритмов машинного обучения для классификации рентгеновских объектов на многоволновых данных"

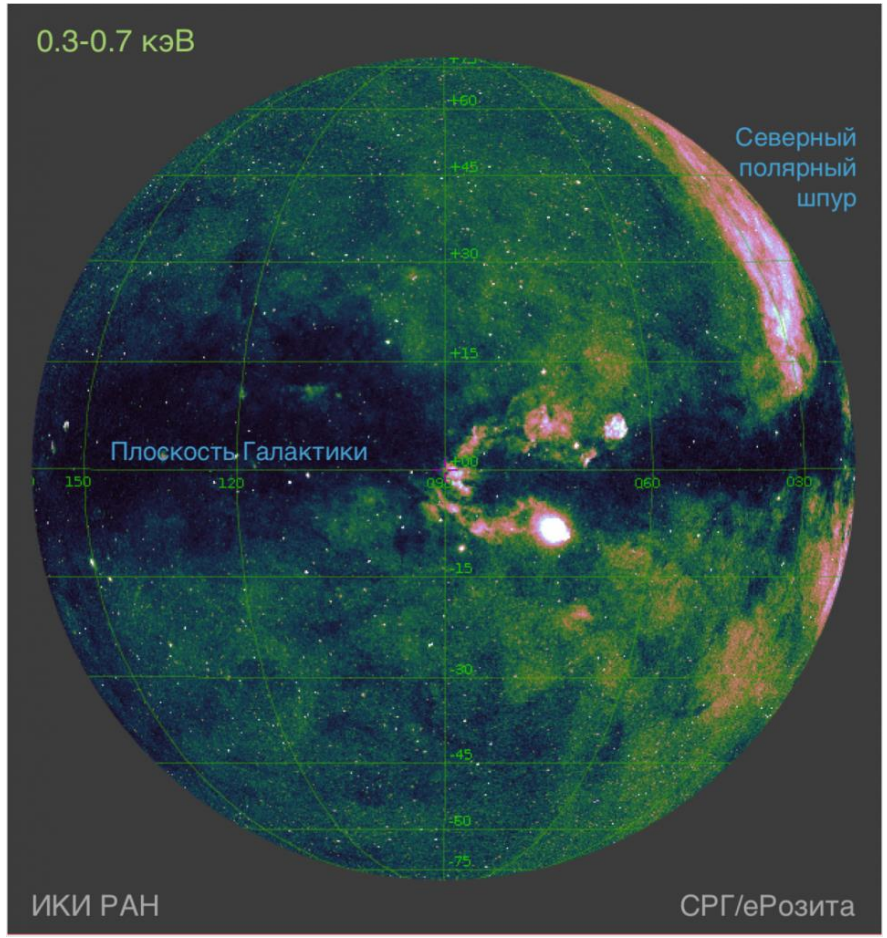
Малышева Надежда
320 группа

Научные руководители:
К.ф.-м.н. Мещеряков Александр Валерьевич
Герасимов Сергей Валерьевич

План:

- Введение
- Актуальность
- Постановка задачи
- Обзорная часть
- Построение решение
- Результаты
- Планы

Введение:

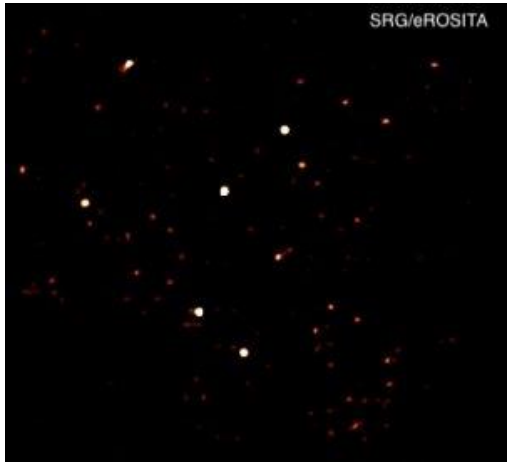


Карта половины всего неба в диапазоне 0.3–0.7 килоэлектрон-вольта, полученная телескопом СРГ/eРозита в ходе первого обзора неба. Изображение: ИКИ РАН

13 июля 2019 года с космодрома Байконур запущена рентгеновская обсерватория SRG. К вечеру 11 июня 2020 года телескоп СРГ/eРозита завершил построение карты, охватывающей всю небесную сферу.

На этих картах зарегистрировано около полумиллиона рентгеновских источников, в том числе звезды, за обработку и анализ изображений половины неба отвечают российские астрофизики.

Таким образом, появляется большое количество неразмеченных данных.

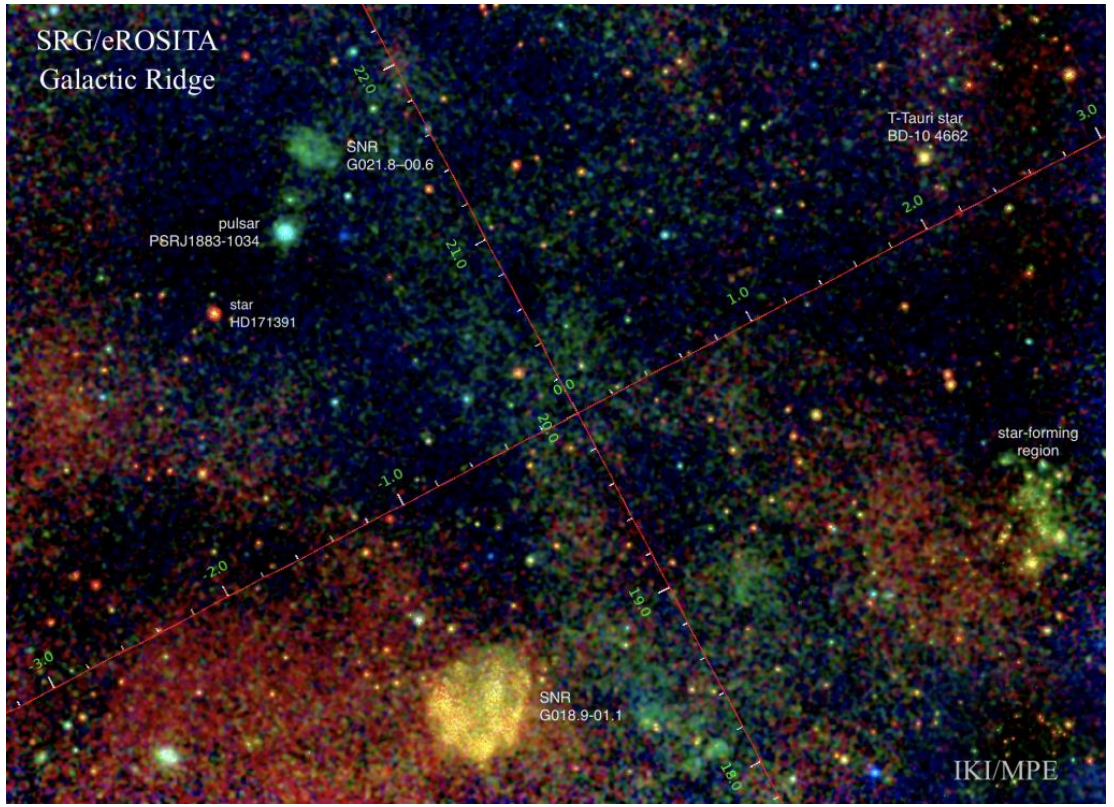


Введение:

Таким образом, появляется большое количество неразмеченных данных.

Задача классификации галактик, квазаров и звезд - одна из самых фундаментальных в астрономии.
Ее решение даст возможность построение новых каталогов рентгеновских объектов.

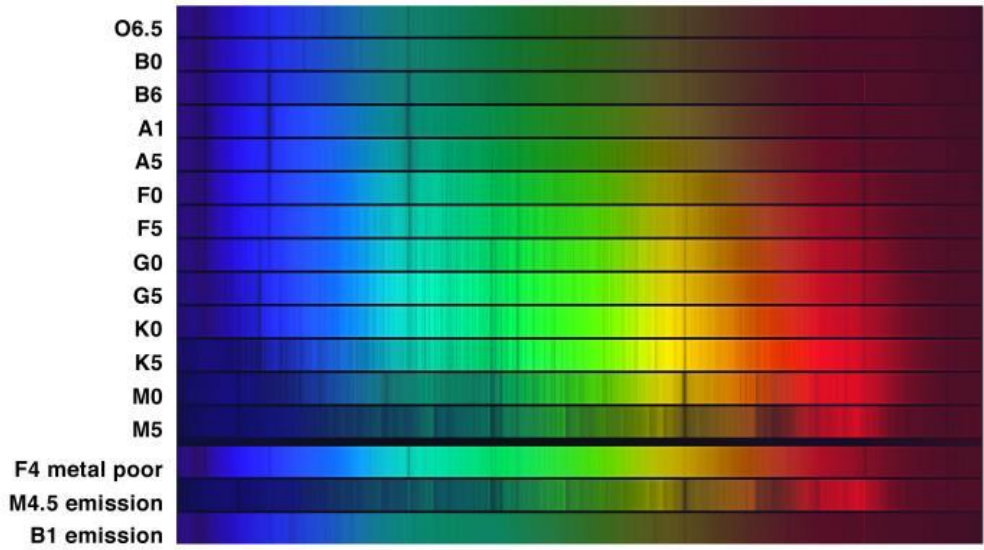
- В частности, классификация и обнаружение рентгеновских двоеных звезд даст возможность исследовать физику экстремального состояния, которую невозможно воссоздать в лабораторных условиях.
- А также исследование активности обычных звезд в рентгеновском диапазоне важно для понимания, около каких объектов возможна жизнь: наличие большой рентгеновской активности говорит, о невозможности существования жизни рядом.



Введение:

Методы классификации можно разделить на **спектроскопические** и **фотометрические**.

Спектрографическая классификация - трудоемкая задача для **слабых оптических объектов**. С другой стороны, классификация может быть основана на данных современных крупных фотометрических обзоров неба, они предоставляют возможность рассмотрения и слабых объектов, и являются более дешевыми с точки зрения наблюдательных ресурсов, чем спектроскопические, но менее точны.



Введение:

Лучи света от звезд, проходя через межзвездную газово-пылевую среду, испытывают поглощение, рассеяние и поляризацию. Для разных длин волн поглощение происходит по-разному. Таким образом происходит искажение света объектов.

$$R_V = \frac{A_V}{E_{B-V}}$$

A_V — величина поглощения
E_{B-V} — изменение показателя цвета B−V (избыток цвета)
$E_{B-V} = (B-V)_{obs} - (B-V)_{real}$

Возможно узнать для звезд с известным спектральным классом. Для далеких объектов поглощение является неизвестным параметром.
Для далеких объектов существенно влияние **межзвездного поглощения** на фотометрические данные.

Введение:

Так же стоит изучать глубокое обучение для табличных данных.
Одними из причин этому могут являться:

- Как и в других областях, можно ожидать повышения производительности за счет архитектур на основе ГНС, особенно для больших наборов данных.
- ГНС используют обратное распространение ошибок данных для управления эффективным обучением от ошибочных сигналов
- Облегчают предобработку данных
- Обучение на потоковых данных
- Эффективно кодируют множество типов данных, такие как **изображения, что можно использовать для повышения точности в текущей задаче.**

На изображениях, кроме данных об объекте могут присутствовать данные об его окружении.

Актуальность:

Таким образом, задача классификации рентгеновских объектов по оптическим данных с помощью методов машинного обучения является актуальной в **области астрофизики**.

Необходимо исследовать применение нейросетевых моделей для классификации на табличных и комбинированных (изображений, таблиц) данных.

Постановка задачи:

Исследовать и разработать модели машинного обучения для классификации рентгеновских звезд, галактик и квазаров по оптическим данным.

- Построение классификатора рентгеновских звезд на основе методов машинного обучения для многоволновых данных
- Построение модели классификации звёзд инвариантной к изменению поглощения и расстояния до объекта.
- Исследование возможности применения нейросетевых методов для данной задачи.
- Исследование и разработка нейросетевых моделей классификации звёзд на основе одновременного использования данных различной природы (изображений, таблиц).

Обзорная часть:

- Обзор данных
- Обзор метрик
- Выбор модели

Обзорная часть:

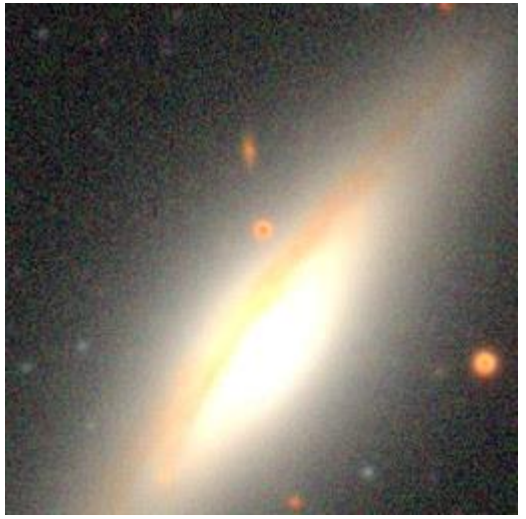
Данные:

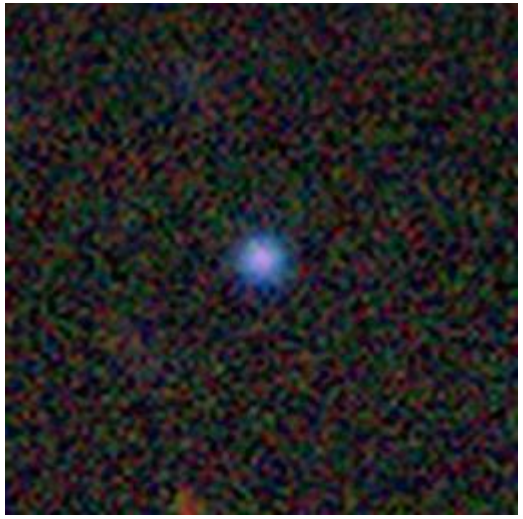
Фотометрические данные могут быть представлены в качестве

- табличных данных:

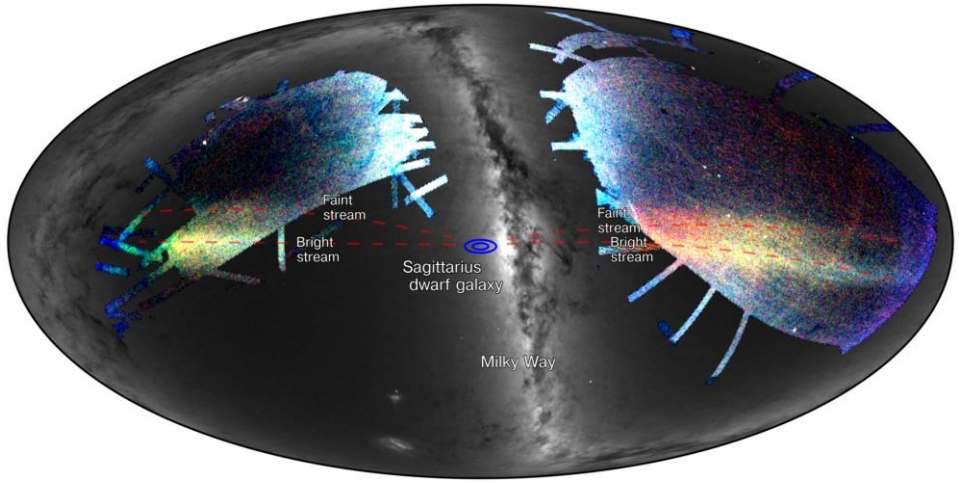
Для SDSS фотометрические измерения в пяти оптических диапазонах: u (λ = 0,355 мкм), g (λ = 0,477 мкм), r (λ = 0,623 мкм), i (λ = 0,762 мкм) и z (λ = 0,913 мкм)) с соответствующими ошибками, представленных в виде величин
psfMag и *cmodelMag (характеризующими поток)*

- И в качестве изображений:



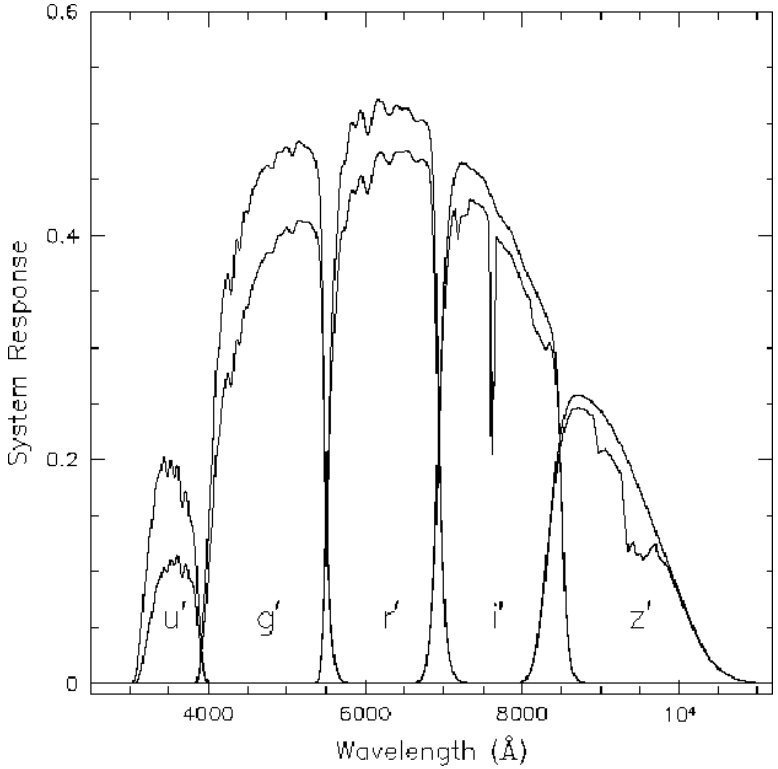


Обзорная часть:



Используются несколько обзоров:

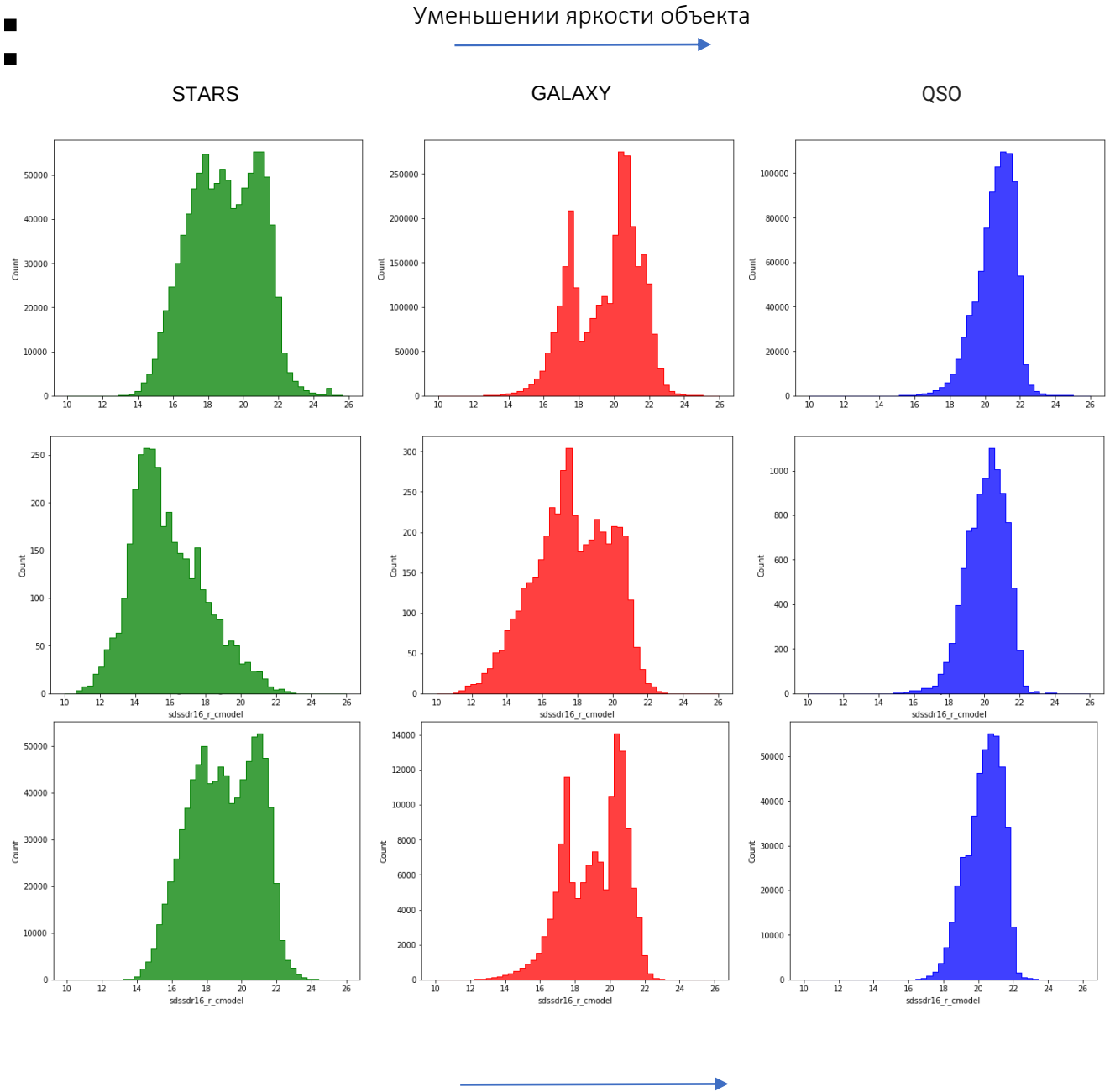
- Обзор SDSS. Фотометрические измерения в пяти оптических диапазонах:
  - u (λ = 0,355 мкм),
  - g (λ = 0,477 мкм),
  - r (λ = 0,623 мкм),
  - i (λ = 0,762 мкм) и
  - z (λ = 0,913 мкм) с соответствующими ошибками, представленных в виде величин *psfMag* и *cmodelMag (характеризующими ядровой и общий поток)* (всего 10 признаков)
- Обзор Pan-STARRS1 в 5 оптических фильтрах (g, r, i, z, y) в виде величин ***psf, kron*** (всего 10 признаков)
- Обзор DESI LIS 3 модельные величины (g, r, z)
- Обзор WISE(инфракрасный обзор всего неба) 2 величины (w1 3,4 мкм, w2 4,6 мкм)



Обзорная часть:

Основные 3 датасета:

- (1)SDSS DR16 (4,6 млн объектов). Только 10 признаков. Больший объем данных (взяты все объекты, имеющие фотометрические данные), присутствуют более слабые объекты.
- (2)Рентгеновские данные (15 тыс объектов). Совмещено несколько обзоров данных (больше различных признаков)
- (3)Только надежные Квазары, Рентгеновские Галактики. Убраны пики по красному смещению. Набор звезд из SDSS DR16.



Обзорная часть (подробное описание используемых наборов данных):

	Кол-во объектов	S	G	Q	Кол-во признаков	SDSS	Pan-STARRS	WISE	DESI (all)	All
(1)	4614588	960363	2789052	865173	10	4614588	-	-	-	-
(2)	20012	2232	3799	7978	42	17336	14210	17403	17329	14009
(3)	1549927	963751	136428	449748	42	1441757	1314912	1445471	1438650	1305157
(4)	1802	42	367	1393	42	1798	1492	1799	1799	1492
(5)	404481	404481	-	-	42	111823	211631	105721	104927	67066

(1) – был получен из SDSS DR16 путем взятия всех объектов, имеющих разметку, удалением дубликатов, строк с недостоверными значениями (-9999) и пропусками. Использовалось 5 фильтров, представленных в виде величин psfMag и cmodelMag.

(2) – был получен путем сопоставления данных SDSS, имеющих разметку класса, а также звезды, определяемые по параллаксу как звезды из каталога Gaia, с данными GAIA DR2 и обзором XMMSSC версии 3XMM DR9, где находится разметка рентгеновских источников.

(3) – получен путем сопоставления размеченных данных SDSS с другими обзорами по ra, dec. Используются только надежные Квазары, Рентгеновские Галактики. Убраны пики по красному смещению. Набор звезд из (1). Добавлены агрегации признаков.

(4) – использовались рентгеновские объекты из обзора Stripe 82X, сопоставленные с данными обзоров SDSS, Pan-STARRS, DESI Legacy Imaging Survey и WISE.

(5) – звезды из APOGEE DR16 StarHorse

Обзор метрик:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Указывает долю верных ответов

confusion matrix (матрица ошибок)

	$y = 1$	$y = 0$
$\hat{y} = 1$	True Positive (TP)	False Positive (FP)
$\hat{y} = 0$	False Negative (FN)	True Negative (TN)

$$precision = \frac{TP}{TP + FP}$$

Показывает, насколько хорошо классификатор определяет истинные положительные результаты (TP), которые являются правильно идентифицированными источниками. Низкая точность для отдельного класса будет указывать на низкую долю положительных идентификаций.

$$recall = \frac{TP}{TP + FN}$$

Указывает, насколько хорошо классификатор сводит к минимуму ложноотрицательные результаты. Низкий уровень отзыва для отдельного класса может указывать на то, что его часто ошибочно классифицируют как другой класс.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

Среднее гармоническое значение точности и полноты и используется в качестве показателя общей производительности.

Случайный лес:

Случайный лес — это метод машинного обучения, состоящий из моделей с древовидной структурой $\{h(x, \Theta_k), k = 1, ...\}$, где $\{\Theta_k\}$ - независимые одинаково распределенные случайные векторы (тренировочные данные). Решение принимается на основе голосования, где каждое дерево дает единичный голос за самый популярный класс на входе x.

Основная схема построения:

1) Повторяется k раз, где k – кол-во деревьев в ансамбле:

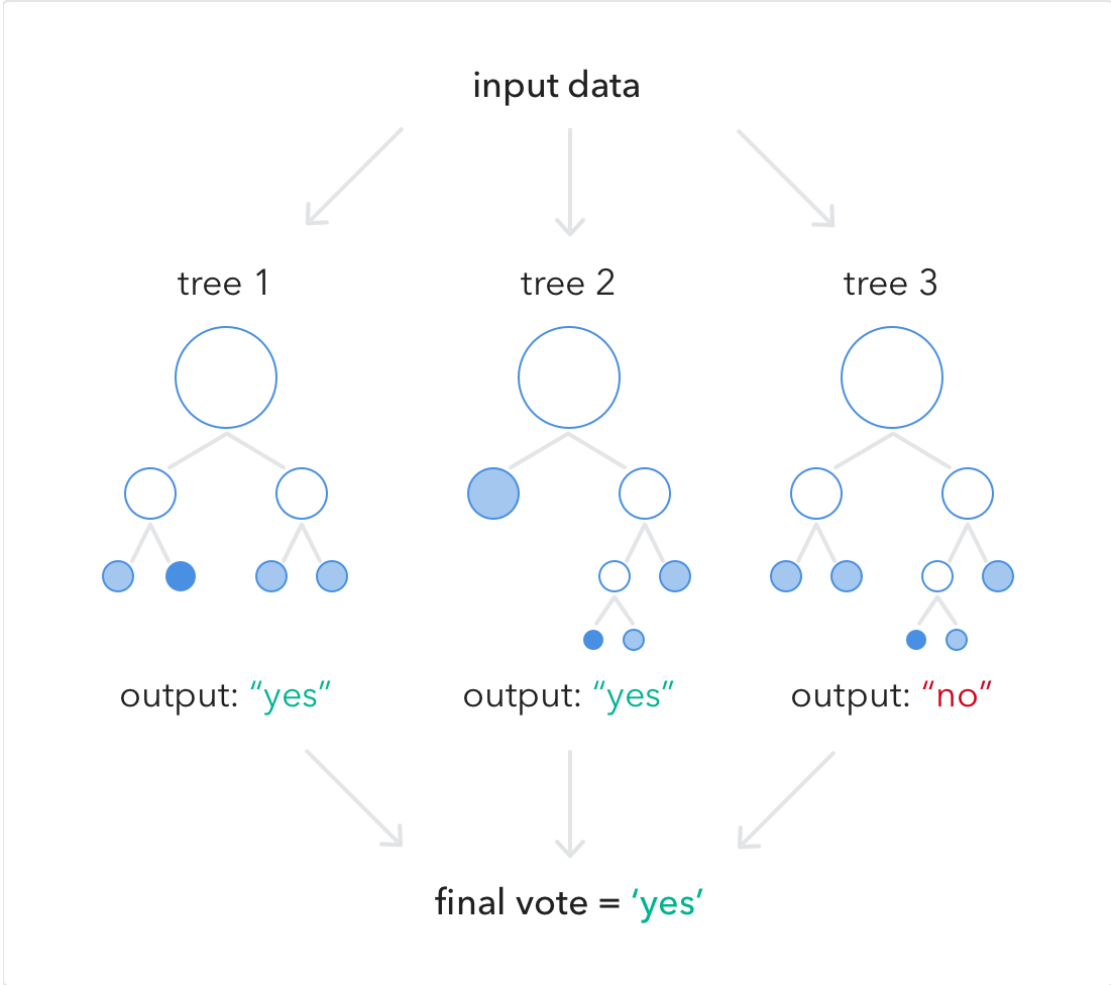
- Сформировать бутстрэп выборку S, размера F по исходной обучающей выборке
- По выбранной S выборке строится дерево решений без ограничения глубины с расщеплением каждой вершины дерева только по фиксированной доле случайно отбираемых признаков.

2) В результате получаем ансамбль из k деревьев решений

3) Предсказание: усреднение предсказания (для задачи регрессии) или голосование (для классификации)

Достоинства:

- Хорошая точность (т. к. деревья в ансамбле слабо коррелируемы)
- Устойчив к выбросу и шуму
- Легкость организации параллельных вычислений
- Прост в подборе гиперпараметров
- Не переобучается



Недостатки:

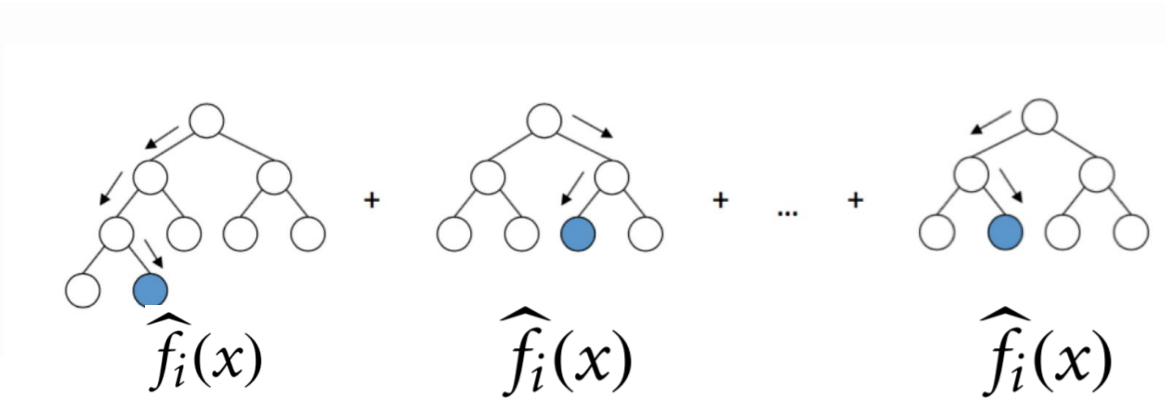
- Плохо работает на разреженных признаках
- Большой размер

Градиентный бустинг:

Градиентный бустинг - это семейство мощных методов машинного обучения, которые показали значительный успех в широком спектре практических приложений.

Основная идея бустинга - последовательно добавлять новые модели в ансамбль. На каждой конкретной итерации новая слабая базовая модель обучается с учетом ошибки всего изученного на данный момент ансамбля.

Оценка представляется в виде: $\widehat{f}(x) = \widehat{f}^{M}(x) = \sum_{i=0}^{M} \widehat{f}_i(x)$



Вход:

- Входные данные $(x, y)_{i=1}^{N}$
- Кол-во итераций M
- Функция потерь $\Psi(y, f)$
- Базовая модель h(x, θ)

Алгоритм:

1: Инициализируем $\widehat{f}_0$ постоянными

2: for t = 1 to M do

3: Вычисляем отрицательный градиент $g_t(x)$

4: Обучаем новую базовую модель $h(x, \theta_t)$

5: Находим лучший размер шага градиентного спуска ρ_t:

$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \Psi\left[y_i, \widehat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)\right]$$

6: Обновляем оценку функции: $\widehat{f}_t \leftarrow \widehat{f}_{t-1} + \rho_t h(x, \theta_t)$

7: end for

Достоинства:

- Мощный метод, который может эффективно фиксировать сложные нелинейные зависимости функций
- Предоставляет множество возможностей для вариаций

Недостатки:

- Идея бустинга обычно плохо применима к построению композиции из достаточно сложных и мощных алгоритмов
- Результаты работы бустинга сложно интерпретируемы, особенно если в композицию входят десятки алгоритмов
- Переобучается
- Плохо параллелится

TabNet:

TabNet – новая высокопроизводительная каноническая архитектура глубокого обучения для табличных данных. TabNet использует последовательные оценки выбора функций, которые следует использовать на каждом этапе принятия решения. Это обеспечивает интерпретируемость и эффективность процесса обучения, поскольку способность к обучению определяется более релевантными функциями

Основная идея: реализовать архитектура глубокого обучения используя древовидную логику.

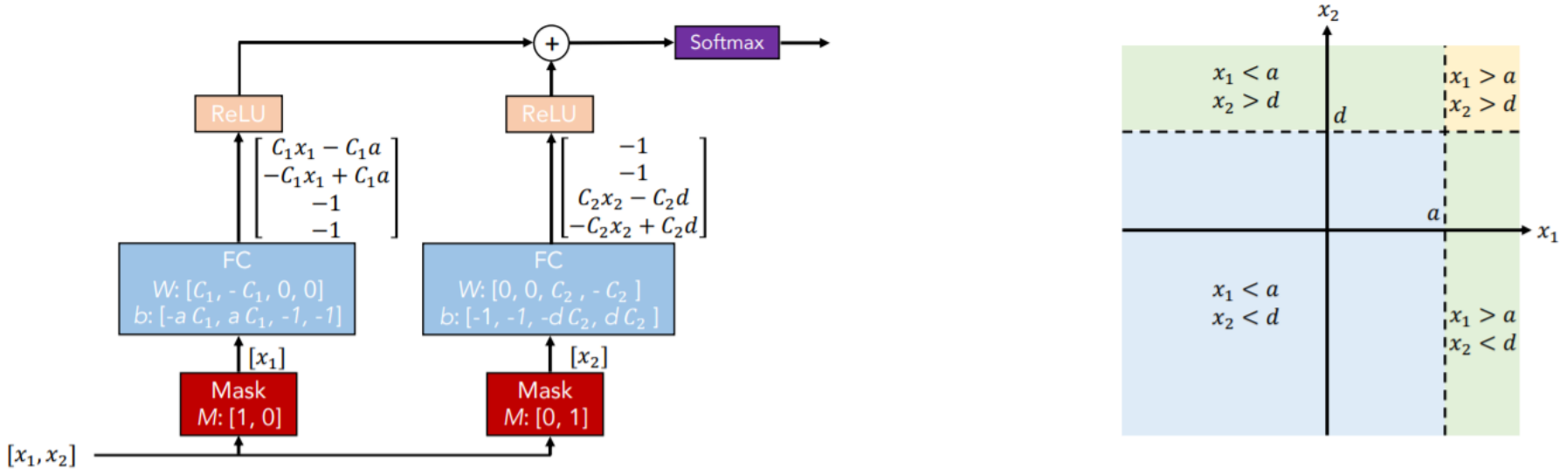
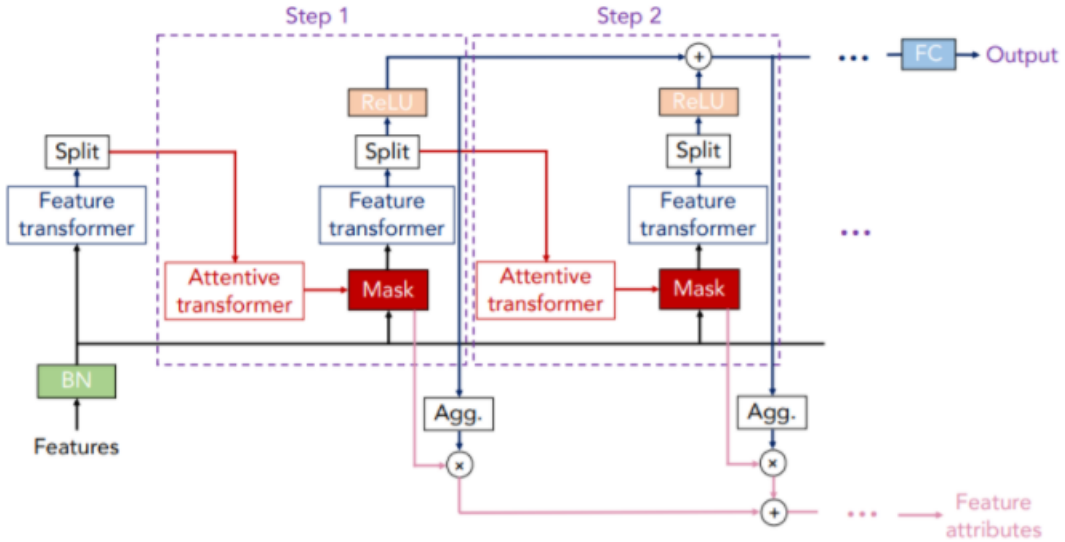
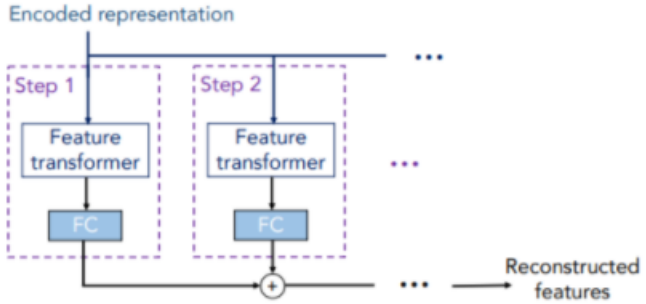


Иллюстрация древовидной классификации решений с использованием обычных блоков ГНС (слева) и соответствующего многообразия решений (справа). Соответствующие объекты выбираются с помощью мультипликативных разреженных масок на входных данных. Выбранные объекты линейно преобразуются, и после добавления смещения (для представления, учета границ) ReLU выполняет выбор области путем обнуления областей, находящихся на отрицательной стороне градиента границы. Агрегация нескольких кластеров основана по аддитивному принципу. По мере увеличения C1 и C2, граница решения становится более резкой из-за функции Softmax (значения классифицирующей логистической функции).

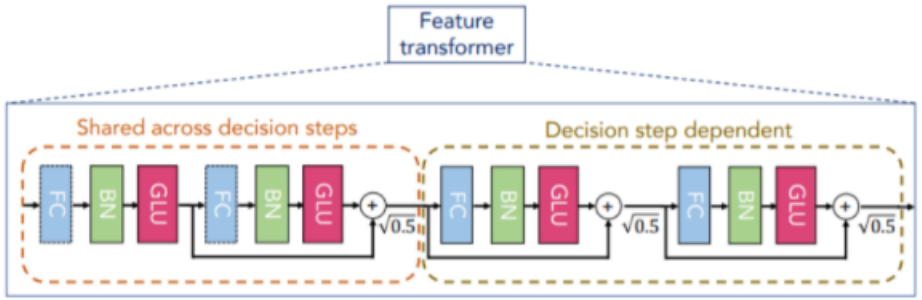
TabNet:



(a) TabNet encoder architecture



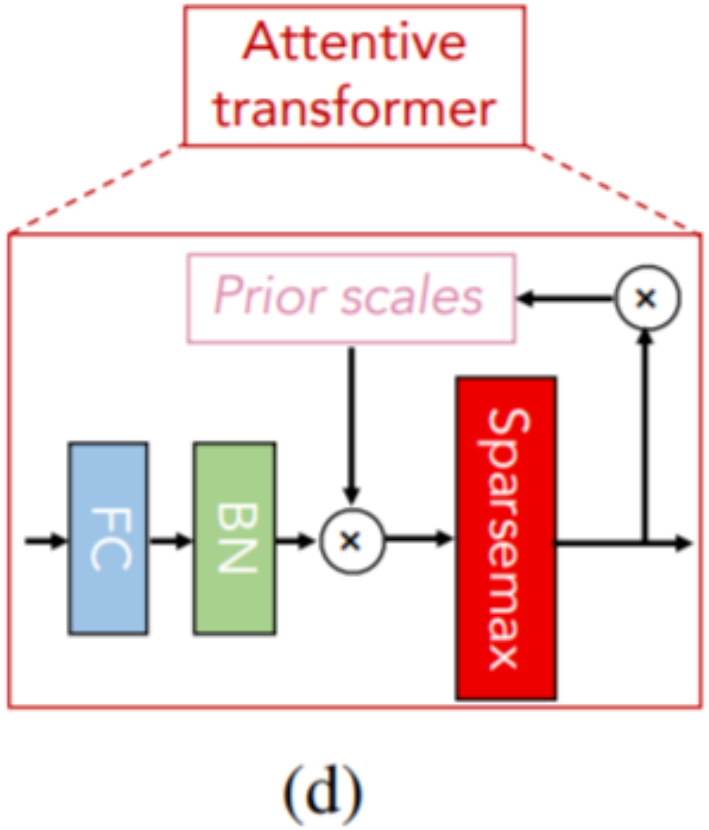
(b) TabNet decoder architecture



(c) Feature transformer

(d) Attentive transformer

TabNet:



(d)

$$f \in R^{(B \times D)} \quad M[i] \in R^{(B \times D)}$$

$$M[i] = sparsemax(P[i-1] \cdot h_i(a[i-1])) \ (1)$$

$$\sum_{j=1}^{D} M[i]_{b,j} = 1$$

$$\mathbf{P[i]} = \prod_{j=1}^{i} (\gamma - \mathbf{M[j]}) \quad (\gamma >= 1)$$

$$\mathbf{P[0]} = 1^{B \times D}$$

$$L_{sparse} = \sum_{i=1}^{N_{steps}} \sum_{b=1}^{B} \sum_{j=1}^{D} \frac{-M_{b,j}[i]}{N_{steps} * B} log(M_{b,j}[i] + \epsilon)$$

TabNet:

Интерпретация важности признаков:

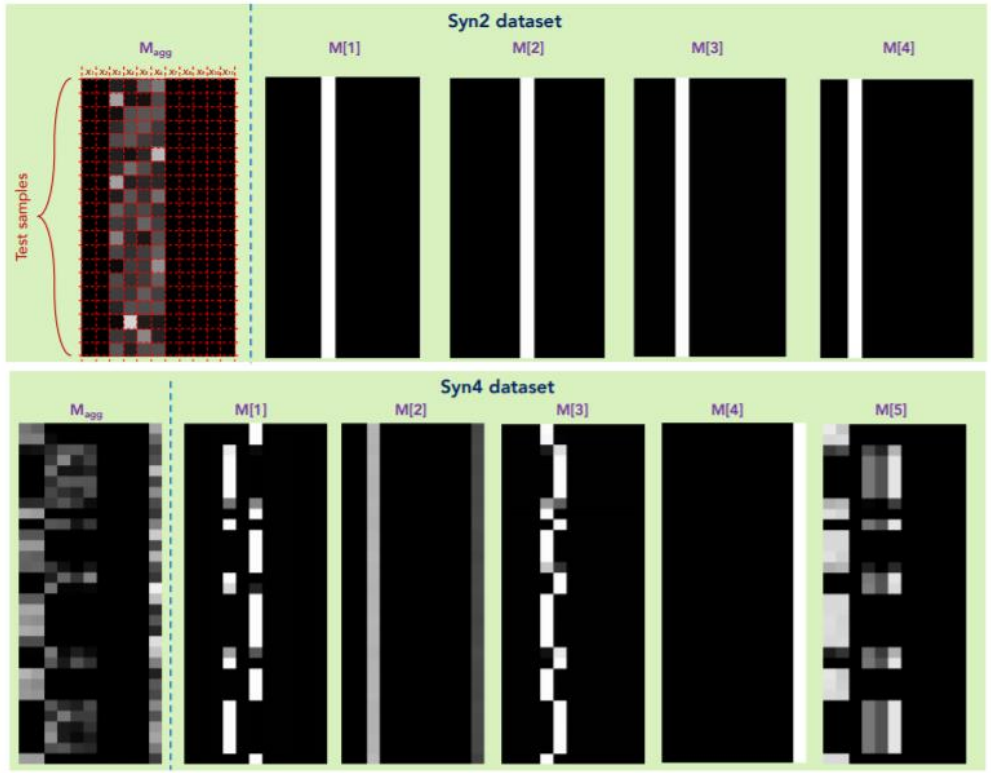
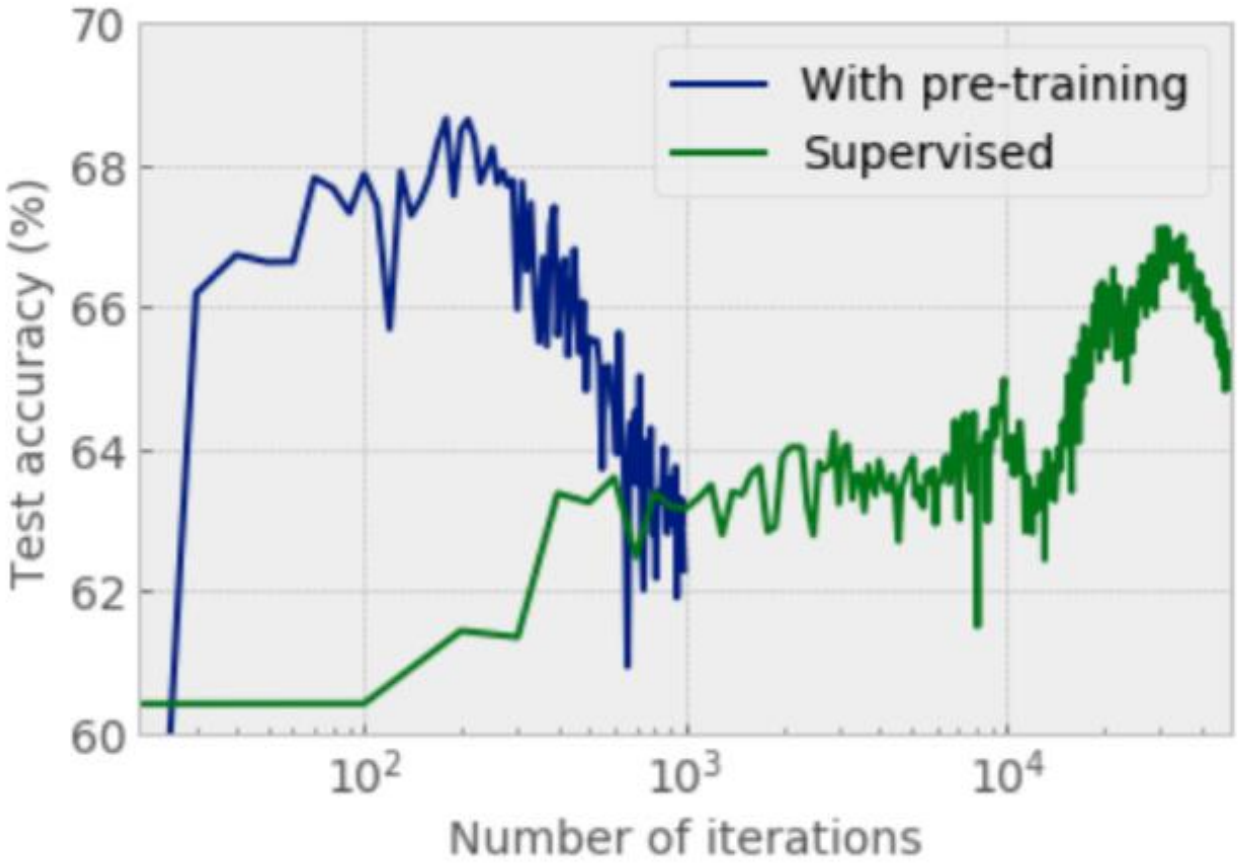


Figure 5: Feature importance masks $\mathbf{M}[\mathbf{i}]$ (that indicate feature selection at i^{th} step) and the aggregate feature importance mask $\mathbf{M_{agg}}$ showing the global instance-wise feature selection, on Syn2 and Syn4 (Chen et al. 2018). Brighter colors show a higher value. E.g. for Syn2, only X_3-X_6 are used.

TabNet:

Возможность предобучения на неразмеченных данных:



TabNet:

Достоинства:

- Хорошая точность для табличных данных
- Интерпретируемость
- Небольшой размер при сохранении модели, за счет небольшого кол-ва весов
- Возможность предобучения
- Малое кол-во гиперпараметров параметров
- Отсутствие необходимости в нормализации данных

Выбор модели:

Для экспериментов будут использоваться TabNet и градиентный бустинг LGBM

Построение решения:

- Построение моделей для классификации рентгеновских объектов
- Построение моделей на данных SDSS
    - Исследование качества модели в зависимости от добавления межзвездного поглощения
    - Сравнения TabNet и LGBM

Построение решения:

Классификация рентгеновских объектов
Для каждой выборки (2) и (3)

Предобработка данных:

- Удаление дубликатов, пропусков
- Нормализация данных
- Разделение на две выборки равномерно по признакам для двукратной кросс-валидации

Построение модели:

- Подсчет точности каждого классификатора на основе **градиентного бустинга** с помощью двукратной кросс валидации и подбора параметров по сетке с использованием hyperopt:

min_child_samples	(1, 50)
colsample_bytree	(0.1, 0.9)
num_leaves	(10, 100)
min_child_weight	(0.001, 0.99)

Построение решения:

Классификация только на SDSS (4,6 млн объектов)

Предобработка данных:

- Удаление дубликатов, пропусков
- Агрегация признаков (добавление линейной комбинации признаков, например: psfMag_i-cModelMag_i, psfMag_r-psfMag_i)
- Разделение равномерно по количеству объектов
- Нормализация данных (только для градиентного бустинга)

Построение модели:

- Использовались модели Pytorch TabNet и градиентный бустинг LGBM
- Подбор параметров для градиентного бустинга параметров по сетке с использованием hyperopt:

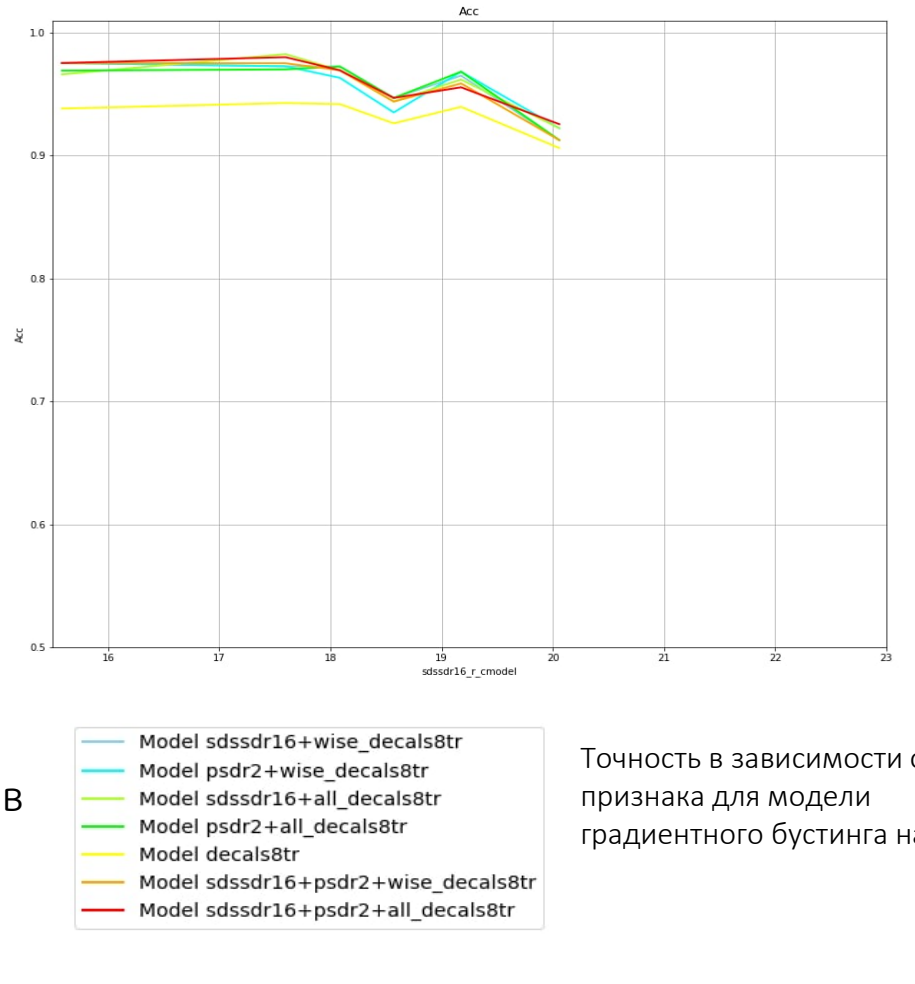
min_child_samples	(1, 50)
colsample_bytree	(0.1, 0.9)
num_leaves	(10, 100)
min_child_weight	(0.001, 0.99)

Результаты:

Data	features	accuracy
SDSS	21	0.964092
DESI LIS+WISE	13	0.958982
Pan-STARRS	20	0.958806
SDSS+DESI LIS+WISE	43	0.972359
SDSS+Pan-STARRS	37	0.96782
Pan-STARRS+DESI LIS+WISE	39	0.968395
All	69	0.973226

Средняя точность в зависимости от используемых обзоров объектов

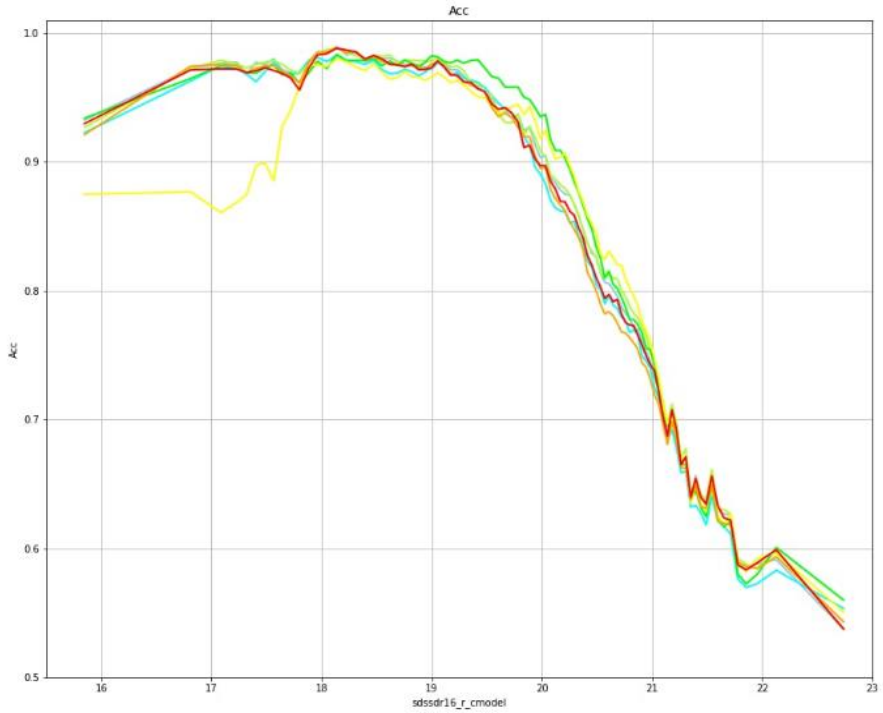
Уменьшении яркости объекта

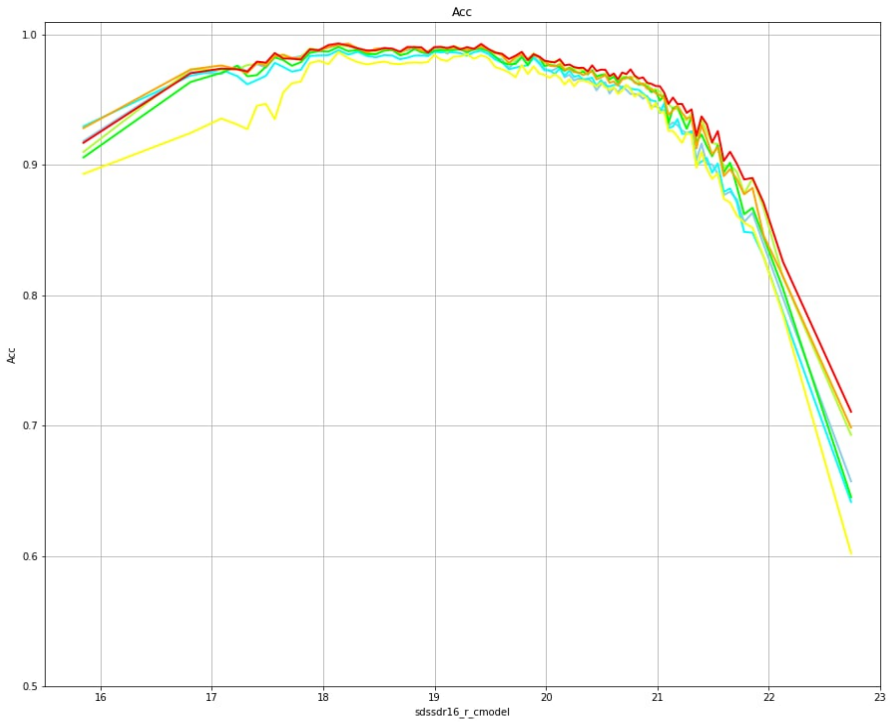


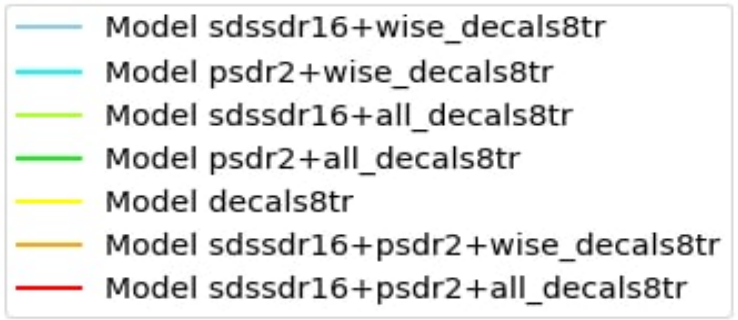
Точность в зависимости от признака для модели градиентного бустинга на (2)

Результаты:

Уменьшении яркости объекта

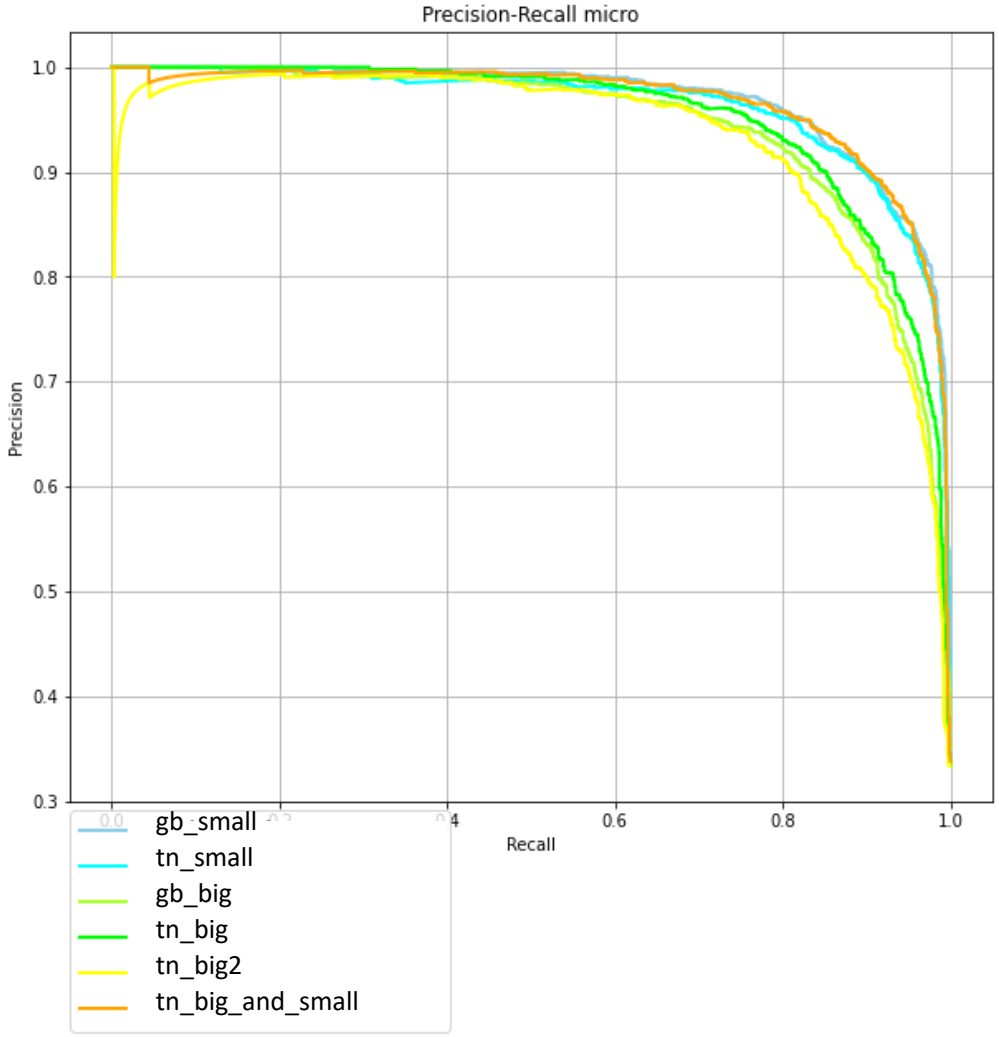






Точность в зависимости от признака для модели градиентного бустинга при тесте на большой выборке:
(a) – модели, обученные на 15тыс рентгеновских объектах;
(b) – модели, обученные на большой выборке (3).

Результаты:



Кривая Precision-Recall на целевой выборке (4) для разных моделей

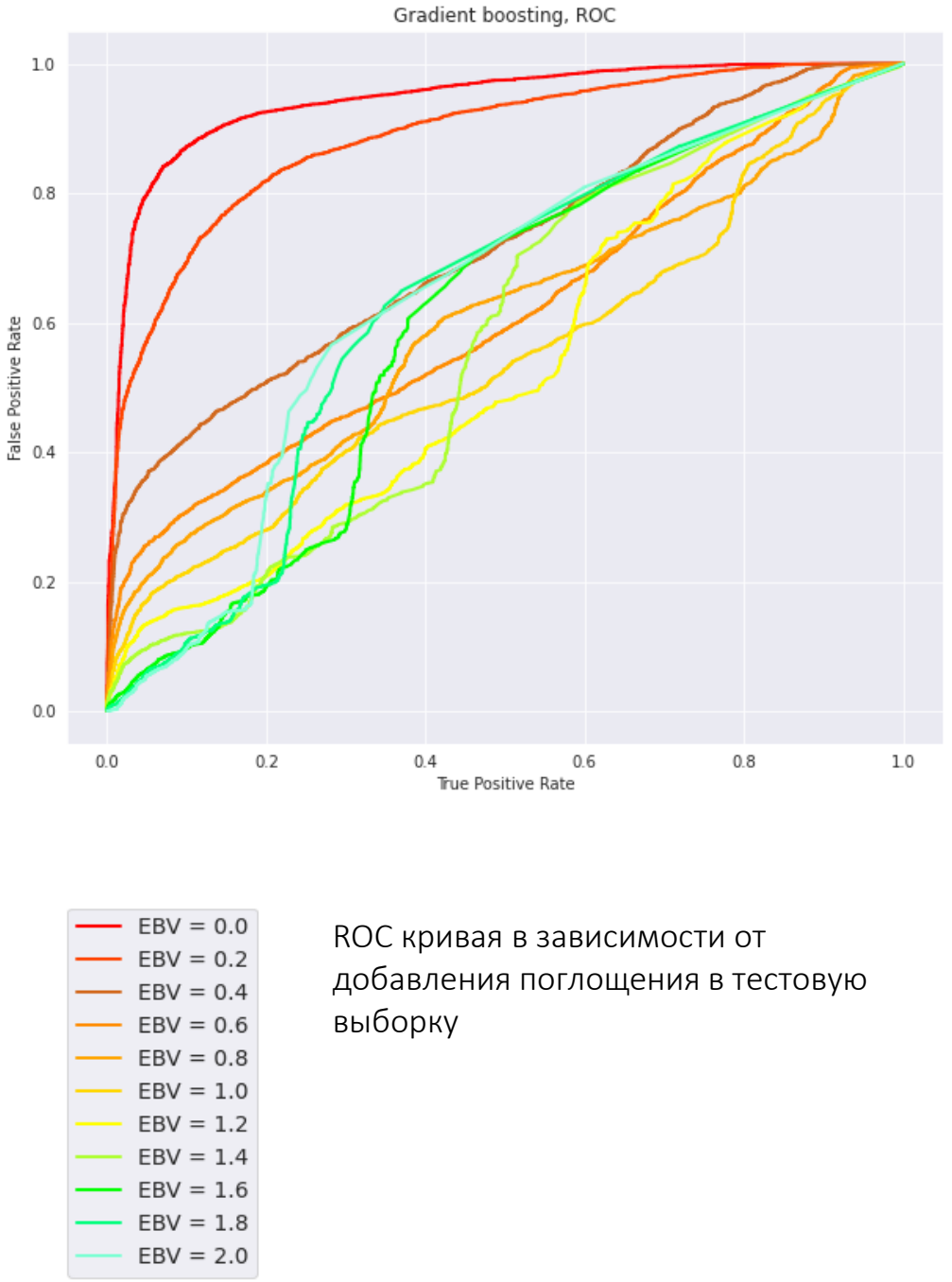
Модель_small – модель, обученная на выборке (2).
Модель_big – модель, обученная на выборке (3).

tn_big2 – TabNet с большим кол-вом параметров на выборке (3).
tn_big_and_small - tn_big2 дообученная на выборке (2) 10 эпох

Результаты:

Исследование влияния поглощения на точность классификации (градиентный бустинг без агрегации признаков на данных SDSS):

При добавлении поглощения заметно существенное ухудшение точности классификации



ROC кривая в зависимости от добавления поглощения в тестовую выборку

Сравнение градиентного бустинга и TabNet:

- TabNet превосходит по точности градиентный бустинг?
- TabNet лучше бустинга отбирает признаки?
- TabNet занимает меньше места?

Сравнение градиентного бустинга и TabNet:

- TabNet превосходит по точности градиентный бустинг?

  (Сравнимые по точности)
- TabNet лучше бустинга отбирает признаки?

  (Бустинг отбирает признаки более качественно)
- TabNet занимает меньше места?

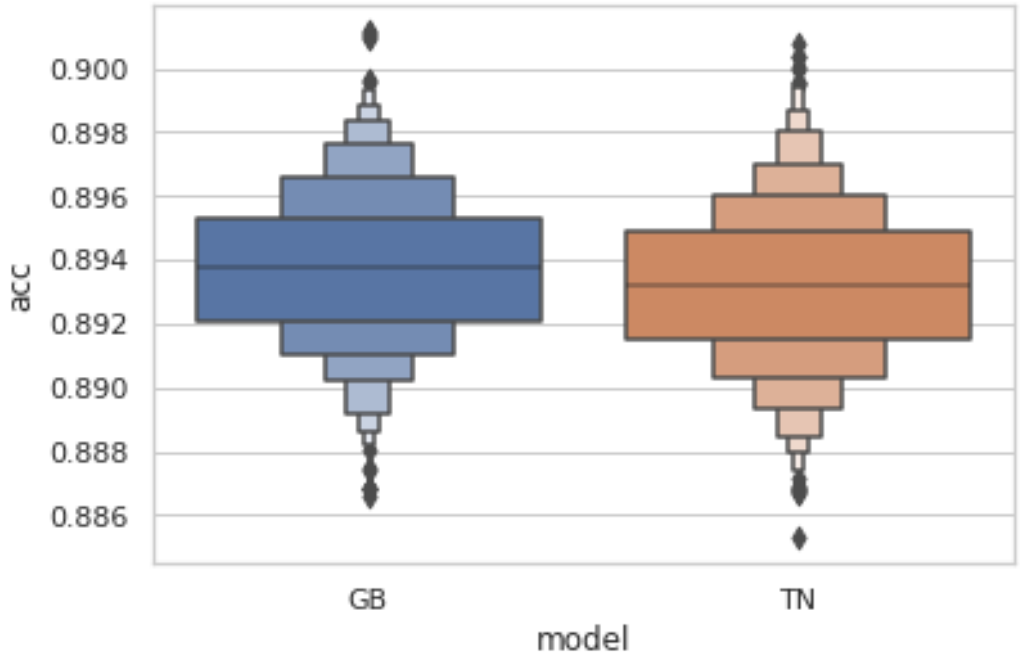
  (Время предсказания на большой тестовой TabNet может быть в разы меньше градиентного бустинга)

Результаты:

Сравнение градиентного бустинга и TabNet:

- Сравнимая точность
- Бустинг выигрывает по скорости обучения

  (пример для модели, обучаемой на 9000 данных :
  TabNet ~ 193 с
  LGBM ~ 3.5 с)



Результаты:

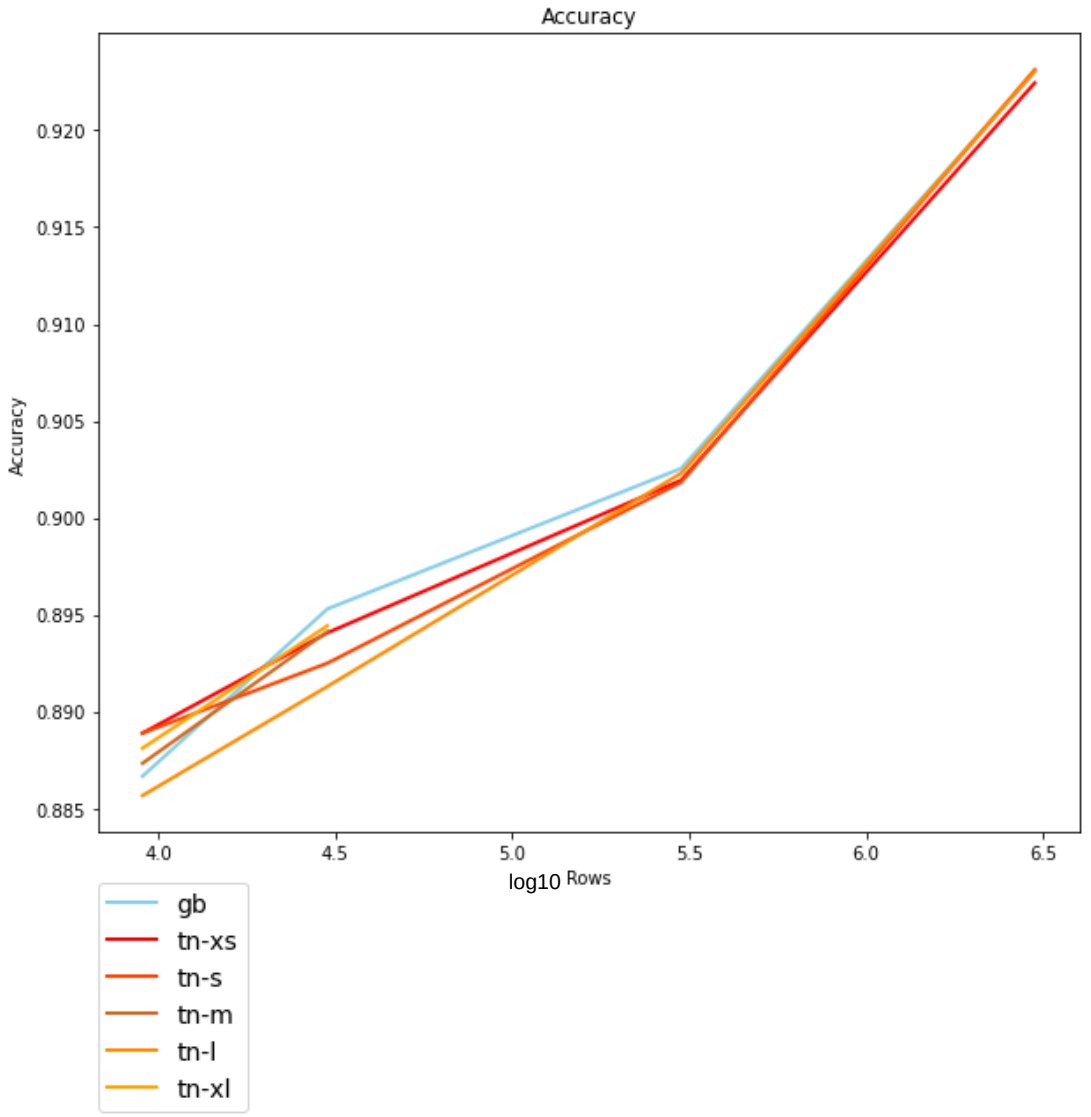
Построены несколько видов моделей TabNet с разным числом параметров

	Число параметров	Nd=Na	λsparse	γ	Nsteps	shared	decision
tn xs	11488	8	0.001	1.5	3	1	2
tn s	32728	16	0.001	1.5	3	2	2
tn m	144328	32	0.001	1.5	3	3	3
tn l	497464	64	0.001	1.7	5	2	2
tn xl	1845496	128	0.001	1.7	5	2	2

Результаты:

	Количество параметров	Размер обучающей выборки 9000	30000	300000	1912769
gb		0.887	0.895	0.903	0.923
tn xs	11488	0.889	0.894	0.902	0.922
tn s	32728	0.889	0.893	0.902	0.923
tn m	144328	0.887	0.894		
tn l	497464	0.886	0.891	0.902	0.923
tn xl	1845496	0.888	0.894		

Точность моделей в зависимости от размера обучающей выборки



Результаты:

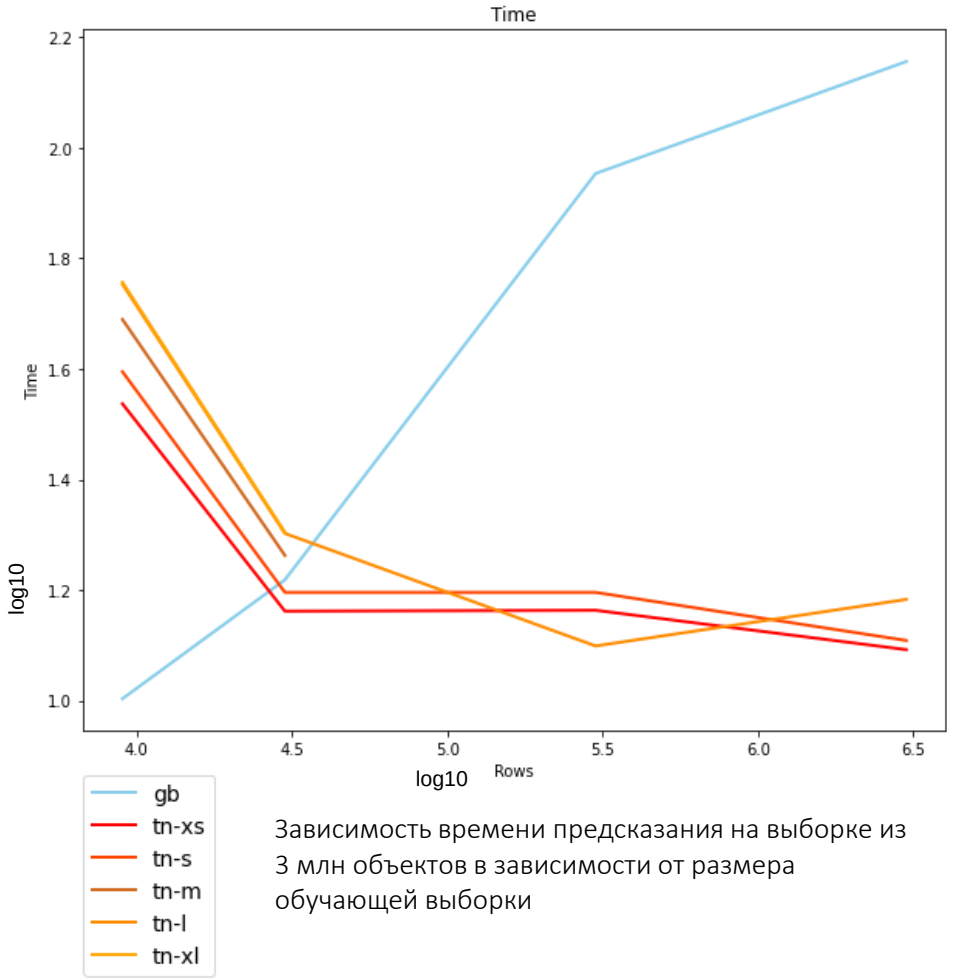
Сравнение градиентного бустинга и TabNet:

- TabNet не увеличивается в размере с увеличением обучающей выборки

Таким образом, TabNet хорош при доступном большом объеме памяти для быстрого предсказания на больших выборках

Размер обучающей выборки	9000	30000	300000	1912769
gb	10.08	16.54	89.78	143.13
tn xs	34.42	14.51	14.57	12.36
tn s	39.33	15.69	15.69	12.85
tn m	48.89	18.30		
tn l	56.62	20.06	12.56	15.24
tn xl	57.07	20.15		

Время выполнения предсказания в секундах на тестовой выборке размером 1912769



Зависимость времени предсказания на выборке из 3 млн объектов в зависимости от размера обучающей выборки

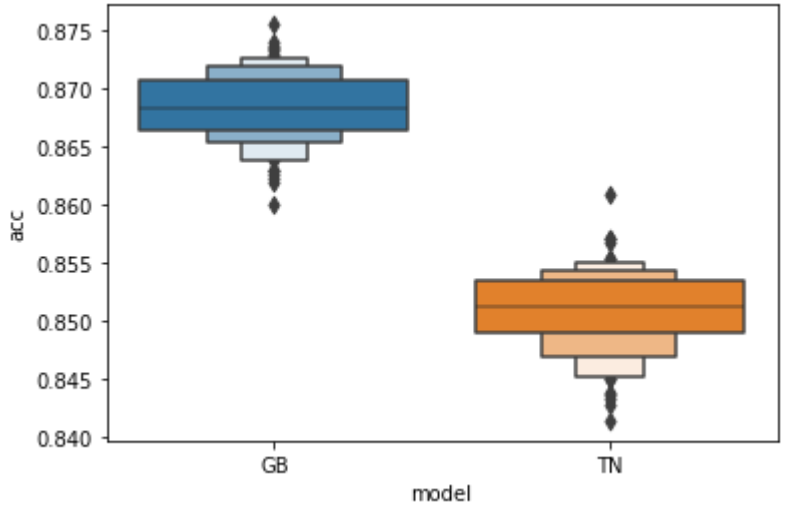
gb (9000)	4.2 MiB
gb(1912769)	45.7 MiB
tn xs	2039MiB
tn xl	2020MiB

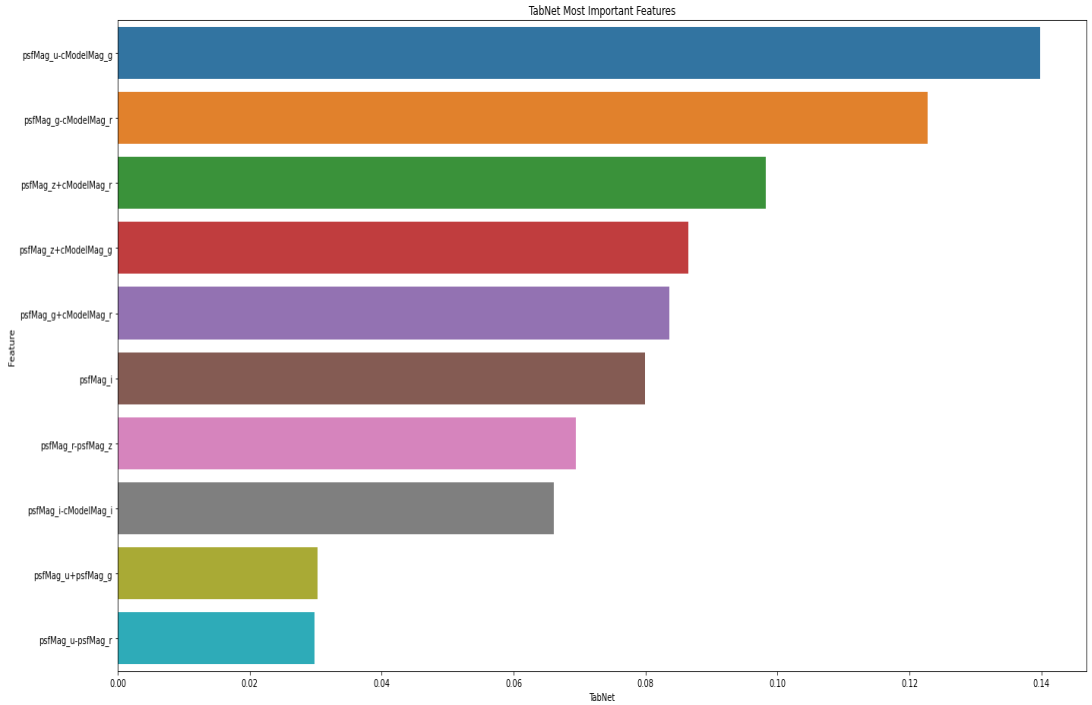
Размер модели в оперативной памяти

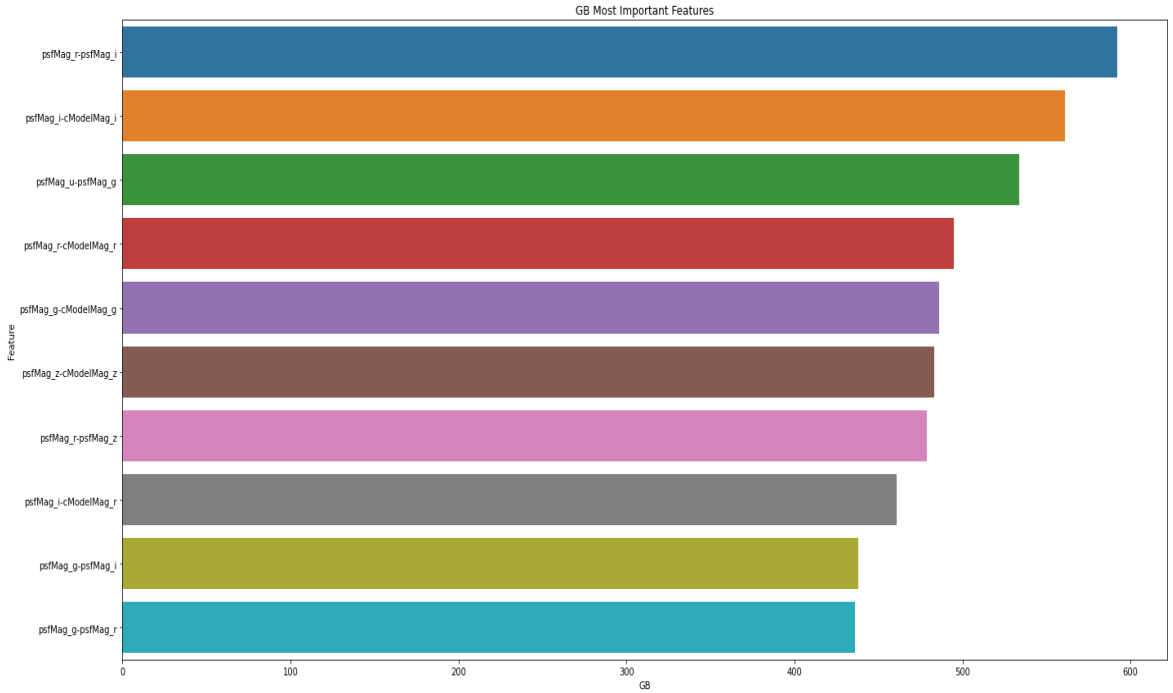
Результаты:

- TabNet хуже отбирает важность признаков

(на первых 3-х важных признаках строился градиентный бусинг и проверялась точность его предсказаний по бутстреп выборке на SDSS данных)







Первые 10 наиболее значимых признака для TabNet и градиентного бустинга

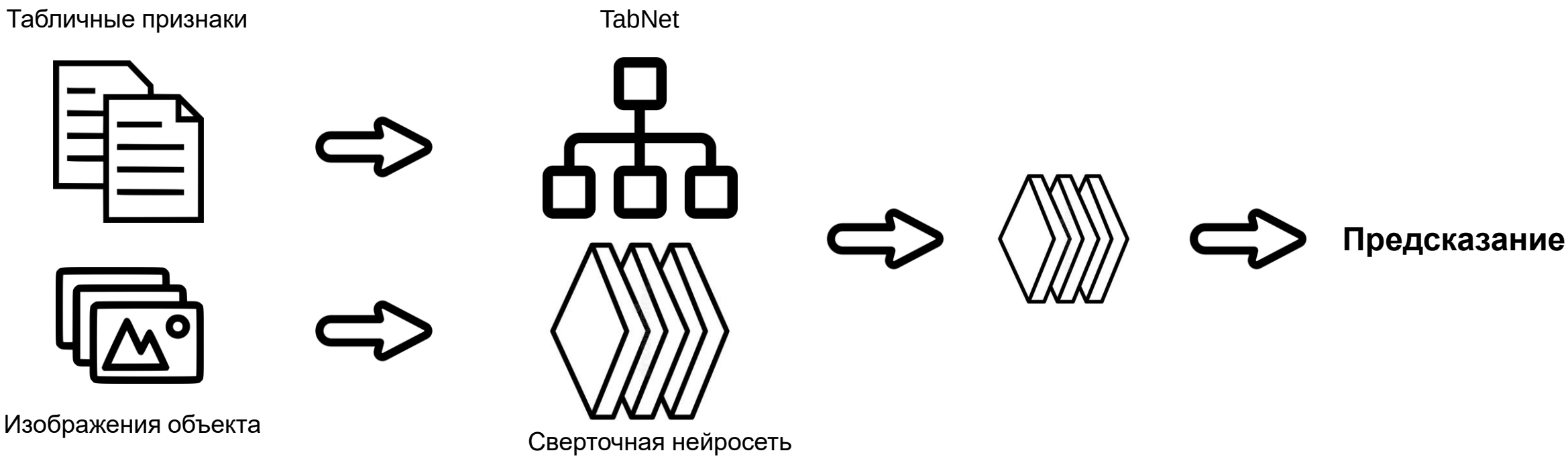
Результаты:

- Исследованы и построены ряд моделей для классификации рентгеновских звезд, галактик и квазаров с использованием различных выборок.
- Проанализировано влияние поглощения на точность модели.
- Построена модель классификации на основе TabNet и проведено ее сравнение с градиентным бустингом по точности, отбору признаков, размеру и скорости предсказания.

Планы:

Так как TabNet – неросеть, то она может обучаться совместно со сверточными нейросетями:

Модель сможет обучаться на большем наборе не коррелируемых признаков



Планы:

- Исследование и разработка нейросетевых моделей классификации звёзд на основе одновременного использования данных различной природы (изображений, таблиц).Исследование предсказания поглощения и расстояния на основе фотометрических данных.
- Построение модели классификации звёзд инвариантной к изменению поглощения и расстояния до объекта.

Спасибо за внимание!