Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра интеллектуальных информационных технологий

Малышева Надежда Валерьевна

Исследование алгоритмов машинного обучения для классификации рентгеновских объектов на многоволновых данных

КУРСОВАЯ РАБОТА

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

Москва, 2021

Аннотация

Классификация космических объектов имеет фундаментальное значение для многих областей астрономических исследований. Учитывая большой объем данных фотометрических съемок, доступных в ближайшем будущем, для этого требуются автоматизированные методы. Были построены несколько моделей машинного обучения (градиентный бустинг и TabNet) для классификации звезд, галактик и квазаров, используя три выборки с фотометрическими табличными данными из каталогов SDSS, Pan-STARRS, DESI Legacy Imaging Survey и WISE. Было исследована возможность применения для данной задачи модель глубокой нейронной сети TabNet, что показало сравнимую точность результата и меньший вес модели при обучении на большом наборе данных.

Оглавление

1. Введение

13 июля 2019 года с космодрома Байконур была запущена рентгеновская обсерватория SRG. Уже в 2020 году телескоп СРГ/eРозита завершил построение карты, охватывающей всю небесную сферу. И обзор неба до сих пор продолжается. Таким образом, появляется большое количество многоволновых данных. Для построения новых каталогов рентгеновских объектов, с целью их подробного исследования, необходима точная разметка класса.

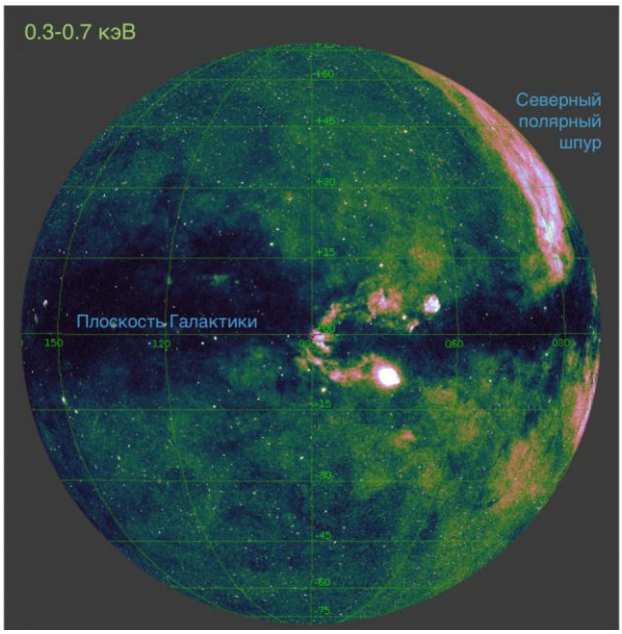


Рисунок 1

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

Задача классификации галактик, квазаров и звезд - одна из фундаментальных в астрономии. Так, например, классификация и обнаружение рентгеновских двойных 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.

Методы классификации можно разделить на *спектроскопические* и *фотометрические*. *Спектроскопический* основан на наблюдении за излучением объекта в различных спектрах. Индивидуальный спектроскопический рисунок позволяет узнать о многих свойствах

космического тела таких, как состав и температура. *Фотометрия* же основана на изучении общего потока излучения. Спектрографическая классификация - трудоемкая задача, так как требует более дорогостоящего оборудования и невозможна для слабых объектов, которые из-за небольшого наблюдаемого размера невозможно разложить на спектры. С другой стороны, современные крупные фотометрические обзоры неба предоставляют возможность рассмотрения всех видимых объектов. Из-за того, что охват длин волн на три порядка меньше, чем у спектроскопии, фотометрия не может уловить те же детали, что и спектры, однако она может уловить общую форму спектра. Сопоставление классов на основе данных затруднительно, поэтому необходимо использовать предсказания моделей машинного обучения. Такой способ классификации более дешевый с точки зрения наблюдательных ресурсов, но менее точный.

При использовании фотометрических данных может возникать проблема искажения оптического потока при поглощении части света, и переизлучении его в другом направлении звездной пылью, расположенной на луче зрения. Это называется межзвездным поглощением. Таким образом, в задачи классификации объектов по многоволновым данным поглощение становится неизвестным параметром, который необходимо учитывать, при рассмотрении областей с сильным межзвездным ослаблением.

Фотометрические данные удобно представлять в виде величины излучаемого объектом общего и центрального потока в различных цветовых спектрах, то есть в виде табличных данных. Самые популярные модели машинного обучения для классификации табличных данных используют ансамблевые деревья решений. И несмотря на то, что это наиболее распространенный тип данных в реализациях ИИ, глубокое обучение для них остается недостаточно изученным. Одними из причин, почему стоит исследовать применение глубоких нейронных сетей для табличных данных, могут являться:

- На примере из других областей можно ожидать повышения производительности за счет архитектур на основе ГНС (глубоких нейронных сетей), особенно для больших наборов данных;
- ГНС используют обратное распространение ошибок данных для управления эффективным обучением от ошибочных сигналов;
- ГНС облегчают предобработку данных;
- ГНС предоставляет возможность обучения на потоковых данных;

- ГНС эффективно кодируют множество типов данных, такие как изображения, что можно использовать для повышения точности при использовании данных различной природы.

Для классификации объектов на основе фотометрических признаков совместное использование табличных данных и изображений может дать значительный прирост в точности, за счет учета не только данных о самом объекте, но и о его окружении (*Рисунок 2*).

Рисунок 2. Пример изображения объекта с его окружением.

Таким образом, задача классификации рентгеновских объектов по оптическим данных с помощью методов машинного обучения является актуальной в области астрофизики.
Необходимо исследовать влияние межзвездного поглощения на точность предсказания, а также рассмотреть применение нейросетевых моделей для классификации на табличных и комбинированных (изображений, таблиц) данных.

2. Постановка задачи

Таким образом задача в рамках курсовой работы состоит в следующем:
Исследовать и построить модели машинного обучения для классификации рентгеновских звезд, галактик и квазаров по оптическим данным.

Исходная задача разбивается на несколько подзадач:

- Построение классификатора рентгеновских объектов 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.

2.1. Формальная постановка задачи:

Пусть X — множество объектов с известными фотометрическими признаками,
Y — конечное множество меток классов для этих объектов. Существует неизвестная целевая зависимость — отображение $y^*: X \to Y$, чьи значения известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.
Требуется построить алгоритм $a: X \to Y$, способный предсказать класс произвольного объекта $x \in X$.

3. Обзорная часть

Для решения поставленной задачи необходимо выбрать данные для обучения и теста, отобрать модели и метрики для проверки качества классификации, а также понять, каким образом происходит искажение данных об объекте межзвездным поглощением.

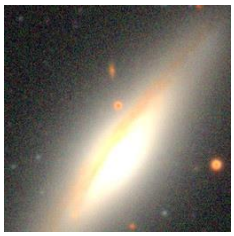
3.1. Данные

Фотометрические данные могут быть представлены в виде изображений и табличных данных (Рисунок 3.

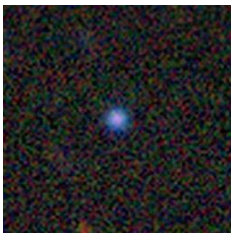
Пример представления фотометрических данных:), полученных путем вычисления величины потока объекта в разных оптических фильтрах.

u	g	r	i	z	Lw1	Lw2
22.81665	22.82024	22.98083	22.51893	22.65033	18.73712	17.50936
23.03668	21.59863	20.28301	19.71039	19.12295	14.63626	14.16189
23.96724	23.01557	22.58311	22.45674	22.35525	17.47329	18.02788
22.93380	21.89849	20.81711	20.21915	19.82858	14.71848	13.78689

(a)



(b)



(c)

Рисунок 3.

Пример представления фотометрических данных:(a) - в виде таблицы, (b), (c) - в виде изображений

Каждый фильтр предназначен для пропускания света определенной длины волны. Фильтры работают, блокируя свет на всех длинах волн, кроме тех, для которых они предназначены. Основными каталогами, предоставляющими оптические данные с размеченными классами объектов, являются SDSS, Pan-STARRS, DESI Legacy Imaging Survey и WISE.

Sloan Digital Sky Survey [1] — это оптический обзор, охватывающий ~ 10 000 deg^2 неба. Данные были получены в обсерватории Апачи-Пойнт с помощью специального 2,5-метрового телескопа и отображены широкоформатной мозаичной ПЗС-камеры. Оптические величины объектов измерялись с помощью пяти широкополосных оптических фильтров u, g, r, i и z 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), представленных в виде величин psfMag и cmodelMag (характеризующими ядровой и общий поток). В основном использовались данные из 16-го выпуска данных (DR16).

Pan-STARRS [2]- это система для получения астрономических изображений с широким полем зрения, разработанная и управляемая Институтом астрономии Гавайского

университета. Pan-STARRS1 (PS1) - это первая завершенная часть Pan-STARRS, которая является основой для выпусков данных DR1 и DR2. В обзоре PS1 использовался 1,8-метровый телескоп и его 1,4-гигапиксельная камера для изображения неба с помощью пяти широкополосных фильтров (g, r, i, z, y) (в виде величин psf, kron). (Рисунок 4).

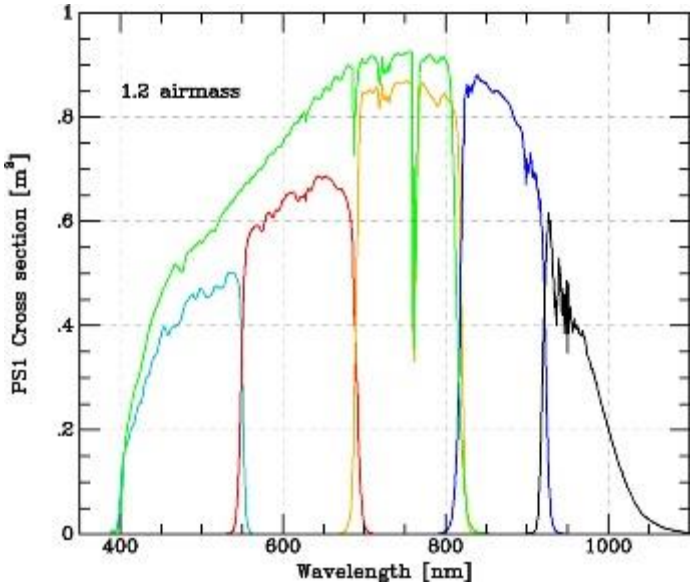


Рисунок 4. Поперечное сечение захвата PS1 в м 2 /e/photon для получения зарегистрированного e^- для падающего фотона для шести полос пропускания Pan-STARRS1, grizy и w для стандартной воздушной массы 1,2.

Исследования DESI Legacy Imaging Survey [3] представляют собой комбинацию трех проектов (Legacy Camera Survey, Пекин-Аризона Sky Survey и Mayall Z-band Legacy Survey), которые совместно снимали ~ 14000 deg^2 внегалактического неба, видимого из северного полушария в трех оптических диапазонах (g, r и z) с помощью телескопов Национальной обсерватории Kitt Peak и Межамериканской обсерватории Cerro Tololo.

WISE [3]обеспечивает фотометрические измерения в четырех инфракрасных диапазонах: w1 (λ = 3,4 мкм), w2(λ = 4,6 мкм), w3 (λ = 12 мкм) и w4 (λ = 22 мкм) с соответствующими ошибками. Использовались только w1 и w2.

На основе данных признаков составленные следующие наборы 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 [4], 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 [5], сопоставленные с данными обзоров SDSS, Pan-STARRS, DESI Legacy Imaging Survey и WISE.

(5) – звезды из APOGEE DR16 StarHorse [6].

3.2. Метрики

Основными метриками для оценки качества многоклассовой классификации являются accuracy, precision и recall. Они основаны на описании точности в терминах ошибок классификации — confusion matrix (матрица ошибок).

Для бинарной классификации ее можно представить следующим образом:

	Ypred = 1	*Ypred = 0*
Ytrue = 1	*True Positive (TP)*	*False Negative (FN)*
Ytrue = 0	*False Positive (FP)*	*True Negative (TN)*

Здесь *Ypred*— это ответ алгоритма, а *Ytrue* — истинная метка класса на этом объекте, положив положительным классом метку 1.

При увеличении количества классов TP, FN, FP, TN вычисляются для каждого класса отдельно, представляя его как положительный, а все остальные метки объединив в отрицательный класс.

Используя эти величины, можно вычислить для каждого класса следующие величины:

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

Указывает долю верных ответов. Ее можно вычислить для всех тестируемых объектов сразу, как отношение верных предсказаний к их общему количеству.

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

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. Используется в качестве показателя общей производительности. При β=1 называется F1-мерой.

Также для описания точности классификации объектов по фотометрическим данным необходимо использовать описанные метрики как функции от признака. Для построения графика такой функции тестовая выборка разбивается на бины по величине, и для каждого бина вычисляется метрика точности.

3.3. Модели

Наиболее популярными моделями для решения задач классификации по табличным данным, имеющим сложные нелинейные закономерности, являются случайный лес и градиентный бустинг.

3.3.1. Случайный лес [7] — это метод машинного обучения, состоящий из моделей с древовидной структурой {h (x, Θ_k), k = 1, ...}, где {Θk} - 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

3.3.2. Градиентный бустинг [8] — 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.

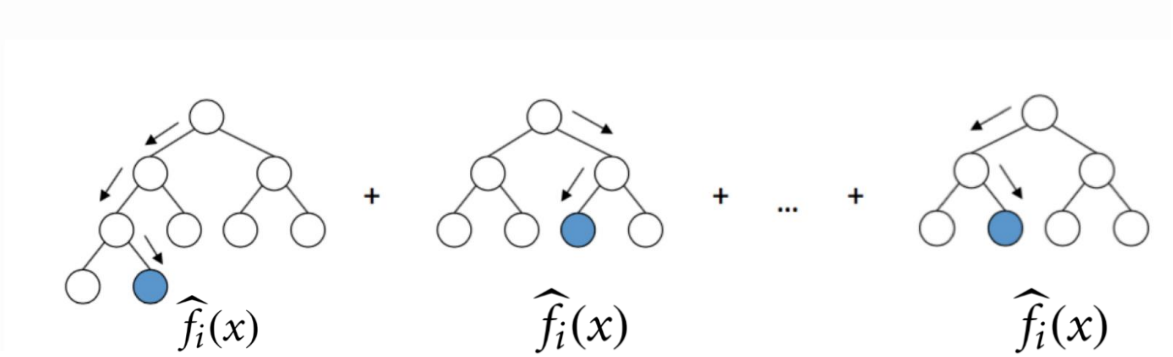


Рисунок 5 Схематическое представление модели градиентного бустинга.

Оценка представляется в виде:

Основная схема построения: $\widehat{f}(x) = \widehat{f}^M(x) = \sum_{i=0}^{M} \widehat{f}_i(x)$

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 постоянными

2) for t = 1 to M do

- Вычисляем отрицательный градиент gt(x)
- Обучаем новую базовую модель h(x, θt)
- Находим лучший размер шага градиентного спуска ρt:

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

- Обновляем оценку функции: ft ← ft−1 + ρth(x, θt)

7) end for

Достоинства:

- Мощный метод, который может эффективно фиксировать сложные нелинейные зависимости функций.
- Более оптимальный в сравнении с случайным лесом по размеру моделей и времени обучения.
- Предоставляет множество возможностей для вариаций (пример: LGBM [9], XGB)

Недостатки:

- Идея бустинга обычно плохо реализуется с композициями из достаточно сложных и мощных алгоритмов.
- Результаты работы бустинга сложно интерпретируемы.
- Переобучается.

3.3.3. TabNet

TabNet зарекомендовала себя, как модель, превосходящая или сравнимая по точности и размеру с градиентным бустингом, на различных типах данных, а также предоставляющая достоверный отбор важных признаков [10]. В связи с этим, интересно проверить ее применимость к данным из астрофизической области. Рассмотрим ее архитектуру.

TabNet – новая высокопроизводительная архитектура глубокого обучения для табличных данных. TabNet использует последовательные оценки выбора признаков объектов, которые следует применять на каждом этапе, несущим вклад в решение. Это обеспечивает интерпретируемость и эффективность процесса обучения, поскольку способность к обучению определяется более значимыми функциями.

Архитектура TabNet схематически может быть представлена следующим образом:

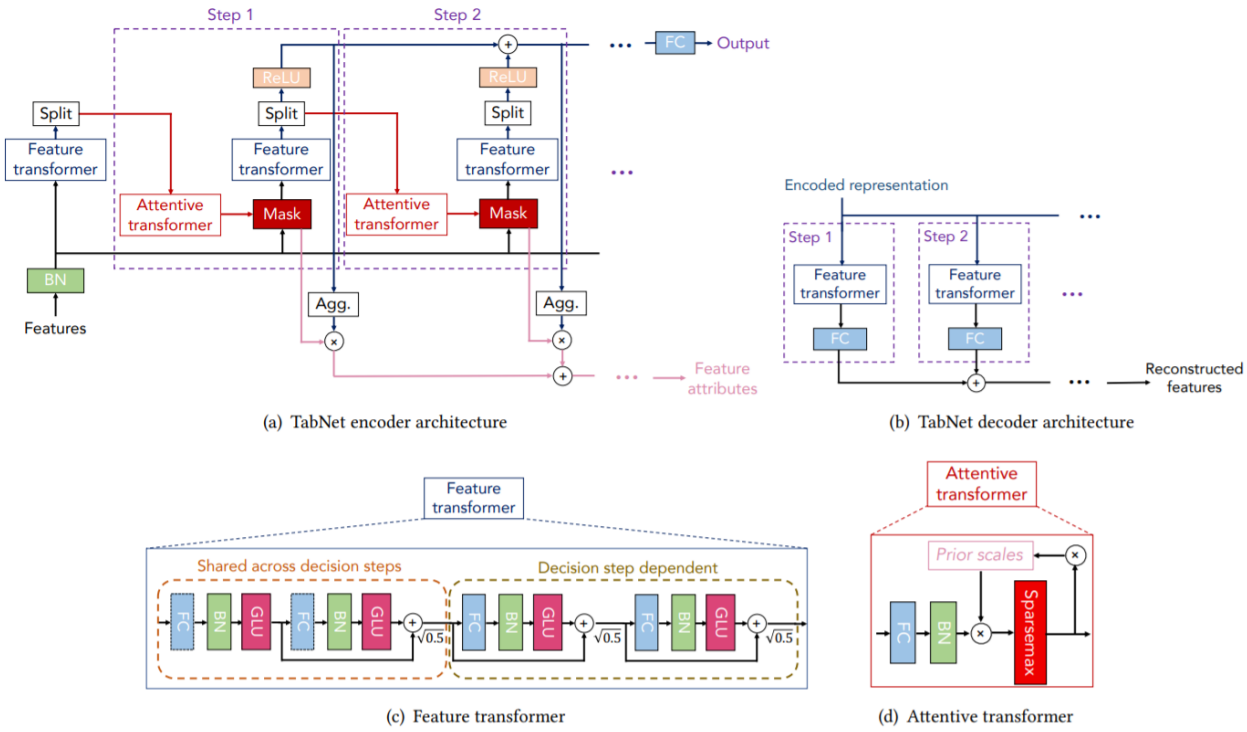


Рисунок 6

(a) Кодер TabNet, который состоит из преобразователя признаков, механизма внимания и маскировки признаков на каждом этапе принятия решения. (b) Декодер TabNet, состоящий из функционального преобразователя на каждом шаге. (c) Пример возможной архитектуры функционального преобразователя – 4-слойная сеть, где 2 блока являются общими для всех шагов принятия решений и 2 зависят от конкретного шага принятия решений. (d) Пример тщательного трансформационного блока, отбирающий на основе информации о частоте использования признака на предыдущих шагах признаки для текущего этапа.

Основная идея: реализовать архитектуру глубокого обучения, используя древовидную логику и механизм внимания.

Для реализации древовидной логики можно использовать разряженные маски на входные блоки для отбора признаков и полносвязые слои для их преобразования. Таким образом, можно представить древовидный классификатор, использую обычные блоки ГНС.

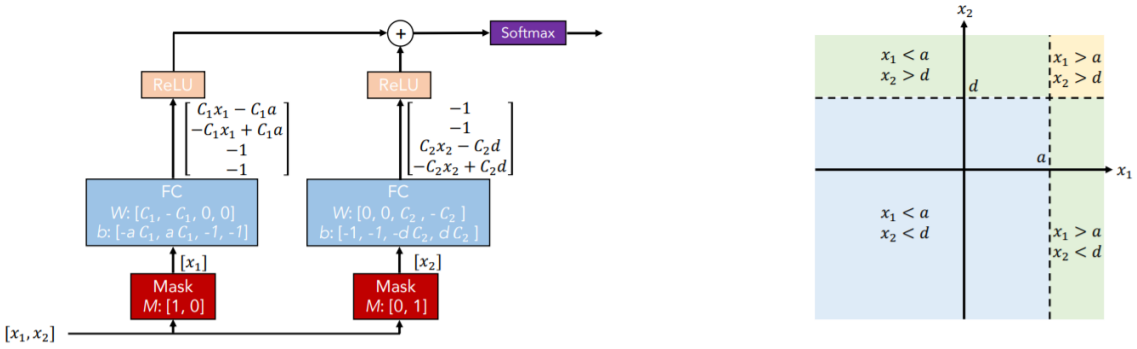


Рисунок 7

Иллюстрация древовидной классификации с использованием обычных блоков ГНС (слева) и соответствующего полученного решения(справа).

Входные данные: $f \in R^{(B \times D)}$

Где B – размер пакета, D – количество признаков.

На i-м шаге:

1) Строим обучаемую маску для отбора признаков M[i] ∈ $R^{(BxD)}$.

   M[i] = *дискретный_преобразователь(P[i - 1]*h_i(a[i - 1]))*

   Где h[i] - обучаемая функция, на Рисунок 6 она представлена в виде однослойного отображения из полносвязного (ПС, Fully-Connected) слоя с пакетной нормализацией (Batch Normalization);

   P[i] - корректирующий показатель, представляющий насколько тот или иной признак был использован ранее: $P[i] = \prod_{j=1}^{i}(\gamma - M[j])$ (γ – параметр свободы, P[0] инициализируются как единицы, или обнуляются, если признаки не используются).

2) Отфильтрованные признаки обрабатываются с помощью функционального преобразователя. Он использует как слои, которые совместно используются на всех этапах, так и слои, зависящие от шага принятия решения. На выходе получаем: $[d[i], a[i]] = fi(M[i] \cdot f)$, где $d[i] \in R^{B \times N_d}$ и $a[i] \in R^{B \times N_a}$

3) Разделяем выход для принятия решения и для передачи следующему этапу для построения маски отбора признаков.

4) Применяя агрегацию признаков, строится полное решение как

$d_{out} = \sum_{i=1}^{Nsteps} ReLU(d[i])$

Для дискретных выходов применяется критерий softmax во время обучения и критерий argmax во время вывода.

Достоинства:

- Показала конкурентную точность на различных наборах данных
- Интерпретируемость
- Небольшой размер за счет использования весов, а не ансамблей из более слабых моделей
- Возможность предобучения
- Малое кол-во гиперпараметров
- Отсутствие необходимости в нормализации данных

Недостатки:

- Долгое обучение
- Не предусмотрена автоматическая агрегация данных

Таким образом, TabNet, как не до конца исследованная на практике модель, может конкурировать с древовидными моделями на табличных многоволновых данных.

3.4. Межзвездное поглощение

Разница в показателях цвета между поглощенными и непоглощенными объектами одного и того же спектрального класса называется избытком цвета и выражается формулой:

$$E_{B-v} = (B - V) - (B - V)_0\,, \tag{1}$$

где E_{B-V}- избыток цвета в звездных величинах;
(B - V) - наблюдаемый показатель цвета;
$(B - V)_0$ – истинный показатель цвета этой звезды.

Зависимость поглощения от длины волны может быть выражена отношением

$$R_V = \frac{A_V}{E_{B-V}}\,, \tag{2}$$

где A_V — величина поглощения;
$E_{B–V}$ — изменение показателя цвета B−V.
В среднем безразмерная величина R_V равна 3,1–3,2.
Отсюда выводится алгоритм определения межзвездного поглощения. Зная спектральный класс наблюдаемого космического тела и сопоставляя с известными заранее нормальными показателями цвета для них, вычисляется избыток цвета. Затем по формуле (2) выводится общее поглощение для данной области неба. Но если нет возможности узнать спектральный класс исследуемых объектов, то величина межзвёздного поглощения остается неизвестной величиной для исходных данный, что может повлиять на точность предсказания класса.
Для каждого фильтра из представленных ранее обзоров существуют коэффициенты (они представлены в статьях [11] [12]), используя которые можно искусственно воссоздать поглощение объекта по формуле:

$$m = m_0 + E_{(B-V)} * coeff(filter, R_V = 3{,}1) \tag{3}$$

Где m_0 – исходная величина, filter – поглощаемый фильтр, coeff – коэффициент поглощения при фиксированной величине R_V, m – результат поглощения.

3.5. Вывод

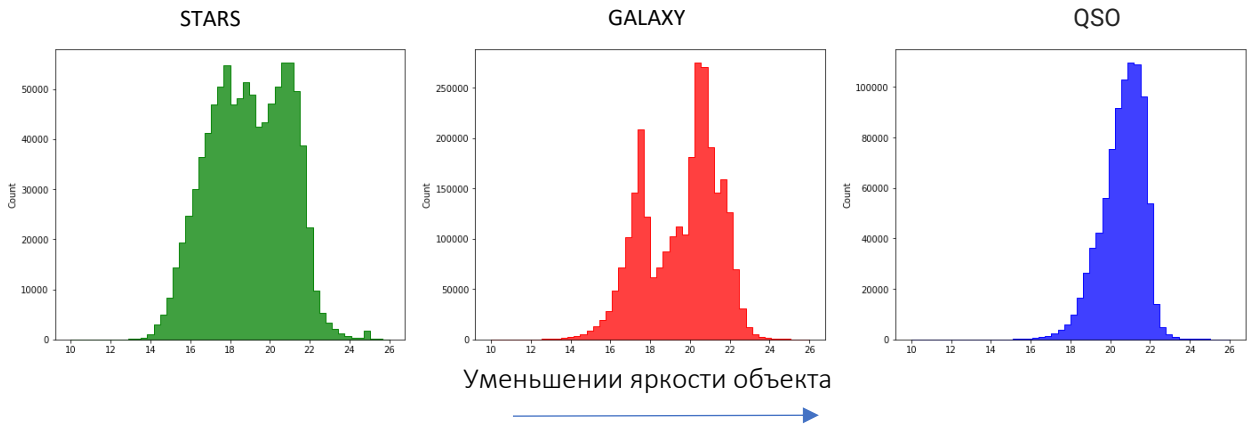
Основными данными для фотометрической классификации представлены в виде данных потока объекта через различные оптические фильтры. Представлены основные методы оценки многоклассовой классификации, которые будут использоваться при классификации космических объектов, и возможные модели для решения поставленной задачи.

Наиболее привлекательной моделью, основанной на деревьях решения является градиентный бустинг, как наиболее современный и оптимизированный метод. Именно он и будет использован как основной для построения классификатора в текущей работе.

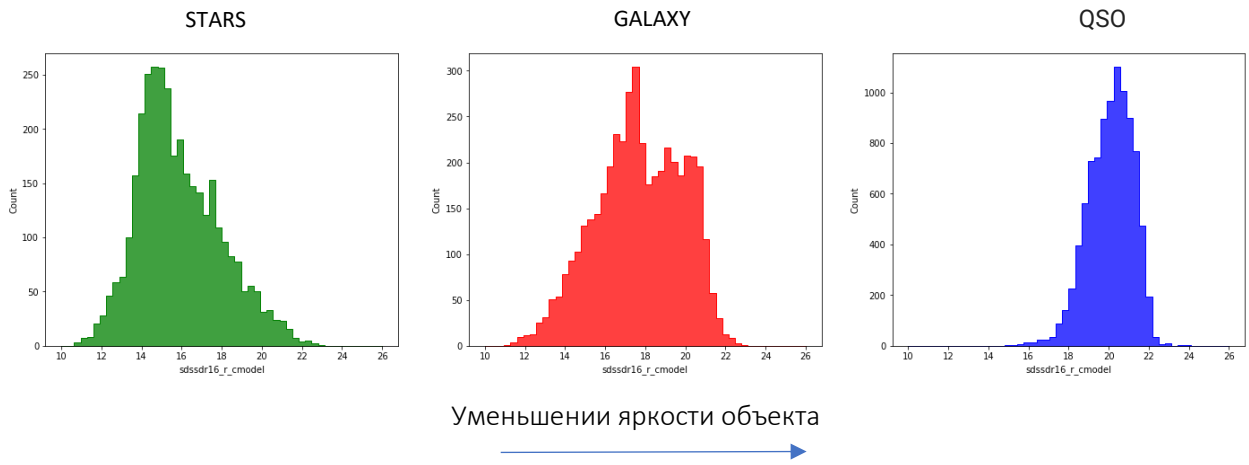
4. Исследование и построение решения задачи

Рассмотрим данные, представленные в обзоре:

- (1) SDSS DR16 (4,6 млн объектов). Только 10 признаков. Представлены все объекты, имеющие фотометрию, без отбора. Главное достоинство - больший объем данных. Также присутствуют слабые объекты величины>20.

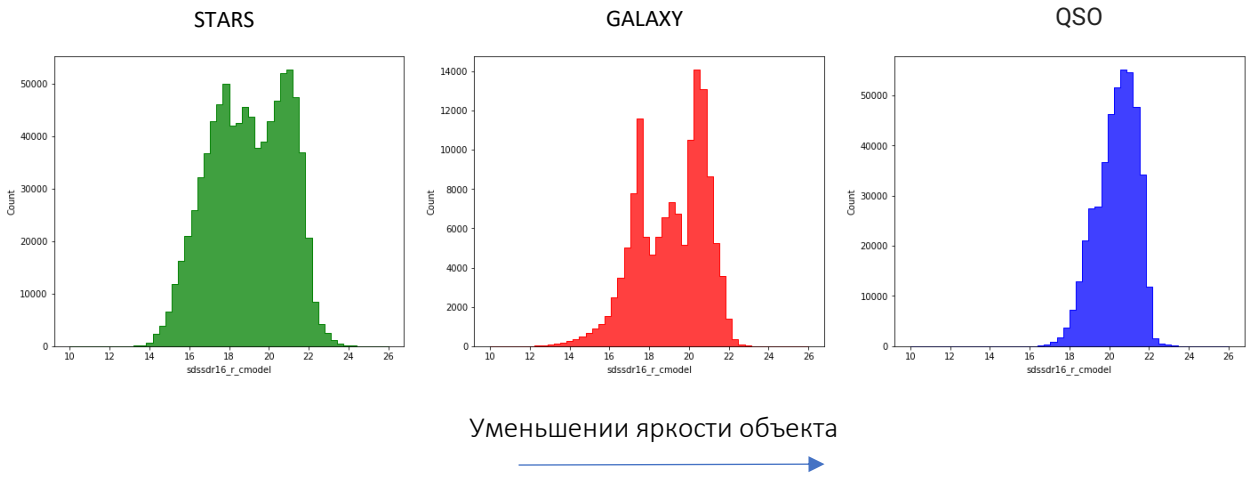


- (2) Рентгеновские данные (15 тыс. объектов, которые можно использовать). Все объектами являются рентгеновскими. В основном представлены только яркими объектами, что является следствием сопоставления с каталогом Gaia (не имеет слабых объектов).

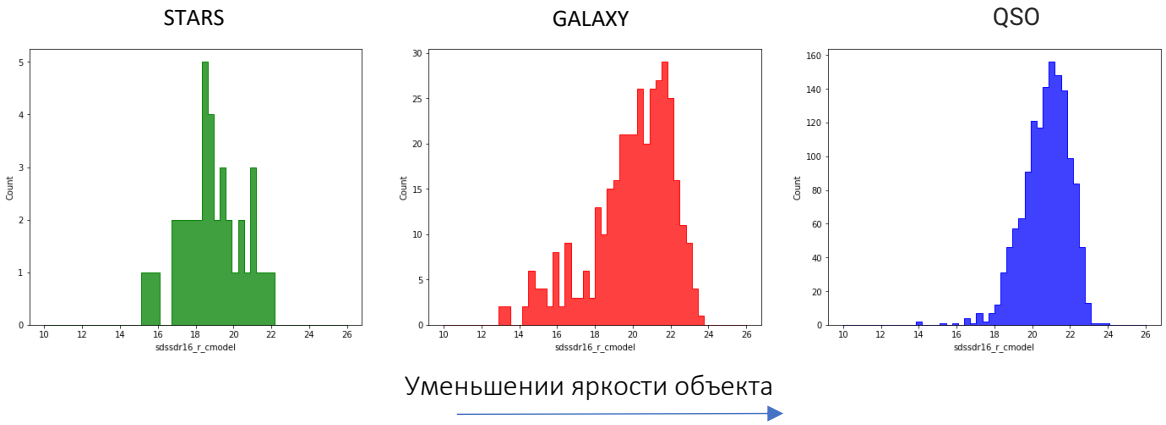


- (3) 1549927 наблюдений. Только надежные Квазары, Рентгеновские Галактики. Убраны пики по красному смещению. Представлено 42 признака из обзоров SDSS,

Pan-STARRS, DESI Legacy Imaging Survey и WISE. Присутствуют слабые объекты величины>20.



- (4) – всего 1802 объектов. Все объекты являются рентгеновскими, наиболее приближены к реальным данным. Поэтому эта выборка лучше всего подходит для теста классификации. Практически нет звезд.



4.1. Построение классификатора рентгеновских объектов на основе методов машинного обучения для многоволновых данных

Так как необходимо классифицировать именно рентгеновские объекты, выборка (2) представляется наиболее привлекательной. Из-за ее небольшого объема проверка точности моделей будет производиться с использованием двукратной валидации:

- Данные разделяются на две выборки равномерно по признаку sdssdr16_r_cmodel.
- На каждой из них *независимо* друг от друга обучаются модели.
- Обученные модели тестируются на неиспользуемой выборке.

Данный метод даст возможность в полной мере оценить точность классификации моделей на данных (2). Далее модели обучаются на всей выборке. Модели, обученные на выборке (2), для определенности будем назвать small.

Также интересно проверить точность моделей, обученных на данных (3). Несмотря на то, что выборка отобрана по всем оптическим объектам, а не только рентгеновским, предполагается, что точность может увеличиться за счет увеличения объема обучающих примеров. Модели так же будут строиться на двукратной валидации и на всей выборке. Модели, обученные на выборке (3), для определенности будем назвать big.

Для проверки точности всех моделей будут использоваться данные (2), (3), (4) – целевая выборка.

В каждой, из используемых выборок, присутствуют признаки из разных обзоров. Для выбора лучших комбинаций данных необходимо обучить несколько моделей, использующие признаки из следующих обзоров: sdss+wise, ps+wise, sdss+decals, ps+decals, decals, sdss+ps+wise, sdss+ps+decals.

Для этого использованы модели градиентного бустинга реализации LightGBM. Использованы набор данных (2), представляющие собой рентгеновские объекты. Было произведено разделения данных на две выборки равномерно по признаку sdssdr16_r_cmodel для двукратной кросс-валидации. Обучалось несколько моделей использующие признаки из разных обзоров: sdss+wise, ps+wise, sdss+decals, ps+decals, decals, sdss+ps+wise, sdss+ps+decals.

В качестве молей машинного обучения используются: преимущественно градиентный бустинг реализации LGBM, а также TabNet, представленные в обзоре.

Для обучения моделей также необходимо произвести подбор гиперпараметров. Оптимальным методом является подбор по сетке с использованием байесовской оптимизации из библиотеки Hyperopt [13] языка Python с проверкой точности на кросс-валидации.

Для градиентного бустинга производится нормализация данных на основе статистических данных из обучающей выборки.

Сетка для градиентного бустинга:

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

Количество ансамблей определялось с помощью early_stopping_rounds для валидацинной выборки.

Сетка для TabNet:

gamma	(1.0, 3.0)
lambda_sparse	(0.001, 0.01)

Остальные параметры подбирались для построения моделей различных по количеству параметров нейросети. Гиперпараметры для моделей представлены в Приложение А.

Дополнительно была обучена модель TabNet на выборке (3) и дообучена на выборке (2), используя возможность нейронных сетей обучаться на потоковых данных (tn_big_and_small).

4.2. Построение модели классификации звёзд инвариантной к изменению поглощения и расстояния до объекта

Для проверки влияния поглощения на классификацию наиболее простым способом является искажение тестируемой выборки с разным поглощением. Самый простой тест может быть произведен на выборке (1) с помощью модели градиентного бустина реализации LightGBM.

Затем такой же эксперимент можно произвести для лучших моделей, построенных на предыдущем этапе. Для добавления поглощения к признакам, каждый фильтр filter из набора u, g, r, i, z изменяется по закону:

$$m = m_0 + E_{(B-V)} * coeff(filter, R_V = 3{,}1) \quad (4)$$

Где m_0 – исходная величина, filter – поглощаемый фильтр, coeff – коэффициент поглощения при фиксированной величине R_V, m – результат поглощения.

Графики ROC-кривой и recall для разных классов в зависимости от применяемого к тестовым данным поглощения E_{B-V} могут показать, насколько модели зависят от межзвездного искажения и для каких объектов могут применяться.

4.3. Исследование возможности применения нейросетевых методов

Для этого эксперимента будут использованы модель градиентного бустинга (как самая оптимальная на данный момент и изученная при применении на данных табличной природы) и ГНС TabNet (новая и малоизученная модель).

Сравнение двух моделей будет происходить в виде поиска ответа на следующие вопросы:

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

Для получения ответов на большинство из этих вопросов, можно использовать простую выборку (1), содержащую наибольшее количество оптических объектов, имеющих разметку. Также для повышения точности классификации необходимо произвести агрегации признаков путем добавления линейных комбинаций с уже существующими (Пример: g_* + u_*, g_* - u_* и т.д.). Таким образом, будут добавлены признаки, являющиеся цветом объекта (g_* - u_* разность фильтров), и признаки, не имеющих физических обоснований (g_* + u_* сумма фильтров), что позволит проверить качество отбора важных признаков у каждой модели.
Обучающие и тестовые выборки должны быть разделены сбалансировано по количеству объектов из каждого класса, чтобы достоверно измерить точность классификации. Для градиентного бустинга производится нормализация данных на основе статистических данных из обучающей выборки.

Для обучения моделей, как и и предыдущих подзадачах, необходимо произвести подбор гиперпараметров. Оптимальным методом является подбор по сетке с использованием байесовской оптимизации из библиотеки Hyperopt языка Python с проверкой точности на кросс-валидации.

Сетка для градиентного бустинга:

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

Количество ансамблей определялось с помощью early_stopping_rounds для валидацинной выборки.

Сетка для TabNet:

gamma	(1.0, 3.0)
lambda_sparse	(0.001, 0.01)

Остальные параметры подбирались для построения моделей различных по количеству параметров нейросети:

	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

Для проверки точности отбора важных признаков будут отбираться 3 самых значимых признака от каждой модели, затем на основе них обучаться модели градиентного бустинга на бутстрэп выборках из того же набора объектов, что покажет относительное преимущество одной из моделей.

Чтобы сравнить занимаемый объем памяти моделей, существует несколько подходов:

- Посчитать количество весов для каждой модели
- Посчитать время предсказания на одинаковой выборке (сколько вычислений происходит для получения результата)

5. Результаты

5.1. Построение классификатора рентгеновских объектов на основе методов машинного обучения для многоволновых данных

Модели, построенные на признаках из различных наборов данных выборки (2) - small, дают сравнимую точность на двукратной кросс-валидации из данных (2). Аналогичные результаты получаются для моделей big при тесте на данных (2).

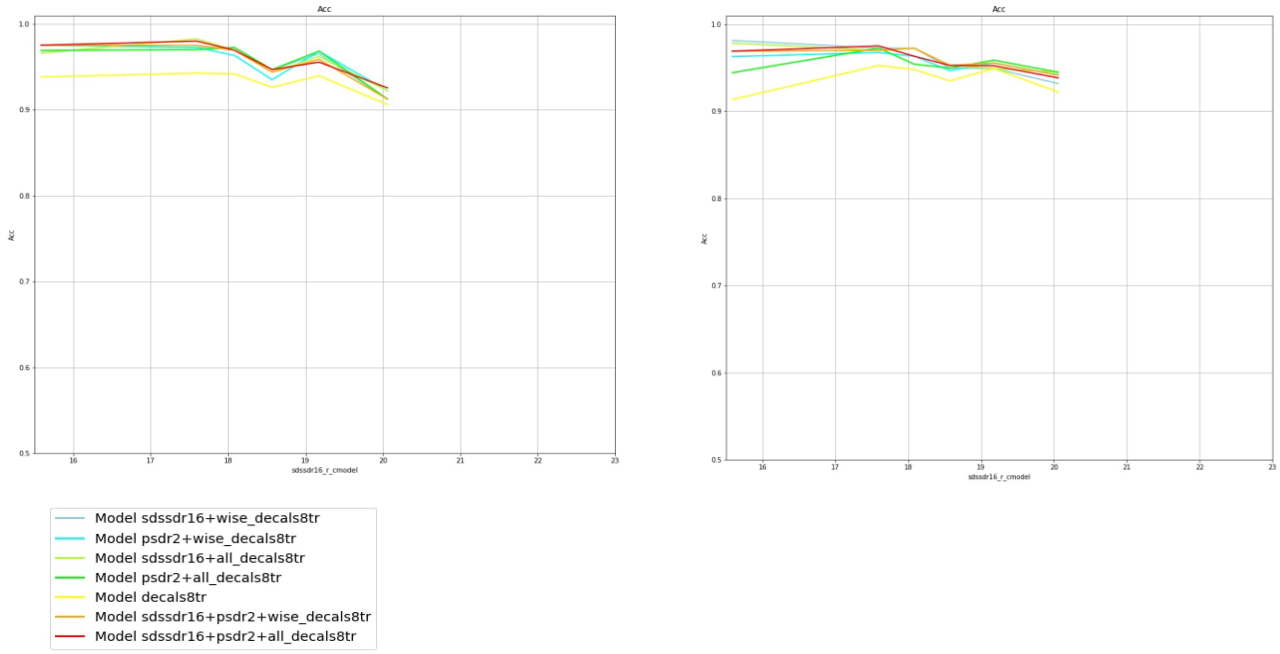


Рисунок 8

Точность моделей (обученных на разных комбинациях признаков) в зависимости от яркости.

(a) – тест small на (2); (b) – тест big на (2).

Тест моделей small и big (на двукратной кросс-валидации) на выборке (3) показал значительное падение точности моделей small на слабых объектах с величиной>19, что происходит из-за отсутствия слабых объектов в обучающей выборке.

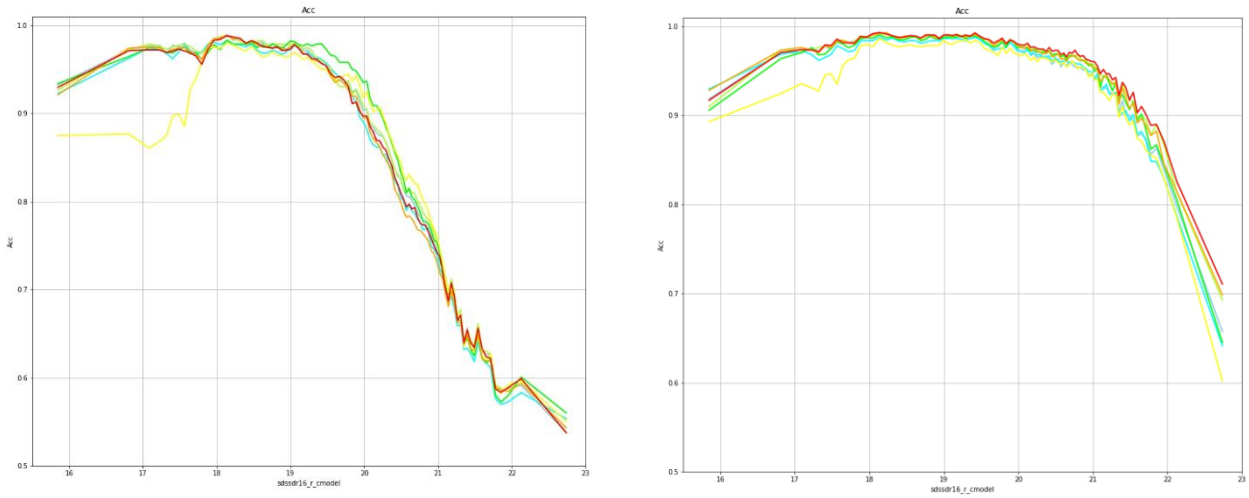


Рисунок 9

Точность моделей в зависимости от признака.

(a) – тест small на (3); (b) – тест big на (3).

По данным результатам, лучшей моделью для классификации обектов, является обученная на комбинации признаков из всех обзоров. Далее будут использованы именно эти модели. (Дополнительные графики в Приложение Б)

Тест на выборке (4) показал следующий результат:

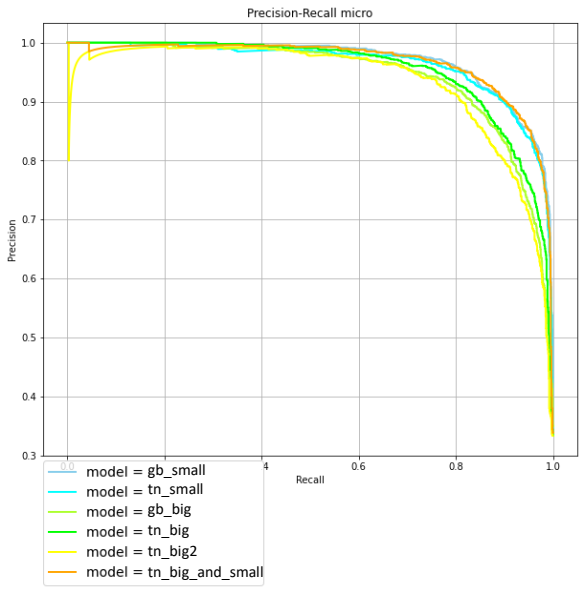


Рисунок 10

Кривая Precision-Recall для разных моделей на данных (4)

Таким образом были построены модели, которые могут при меняться для рентгеновских данных.

5.2. Построение модели классификации звёзд инвариантной к изменению поглощения и расстояния до объекта

При добавлении поглощения в тестовую выборку (для данных (1)) заметно существенное ухудшение точности классификации. При $E_{B-V} > 0.2$ классификатор дает случайный ответ.

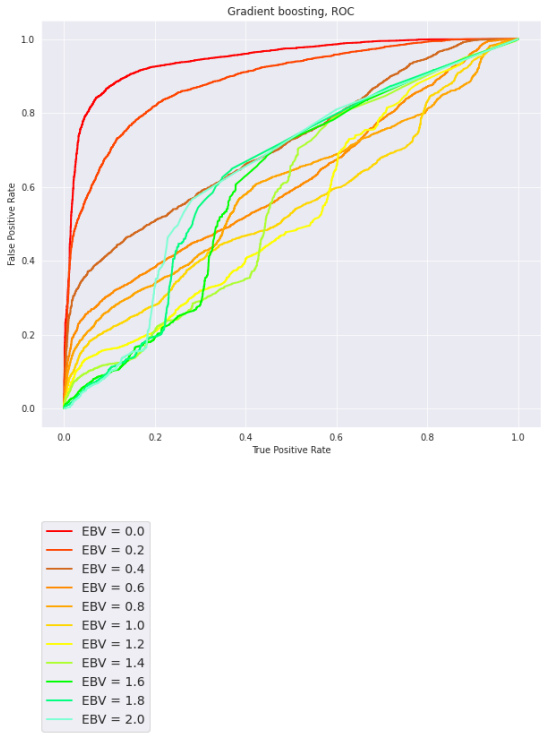


Рисунок 11

ROC-кривые, построенные на тесте для моделей градиентного бустинга на данных (1) при искусственном добавлении поглощения EBV по формуле (3)

5.3. Сравнение TabNet и градиентного бустинга

5.3.1. TabNet превосходит по точности градиентный бустинг?

Проверка точности показала сравнимые результаты моделей TabNet и LightGBM (Рисунок 12

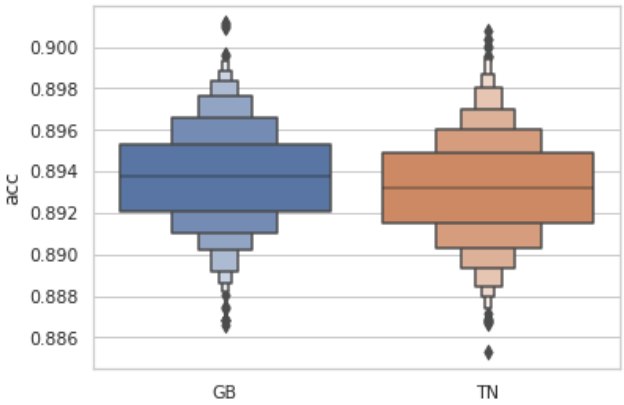


Рисунок 12

Точность моделей градиентного бустинга и TabNet
Точность моделей градиентного бустинга и TabNet).

Также при сравнении моделей, полученных на предыдущих этапах для построения классификатора рентгеновских тел, TabNet дает похожие результаты при большом времени обучения:

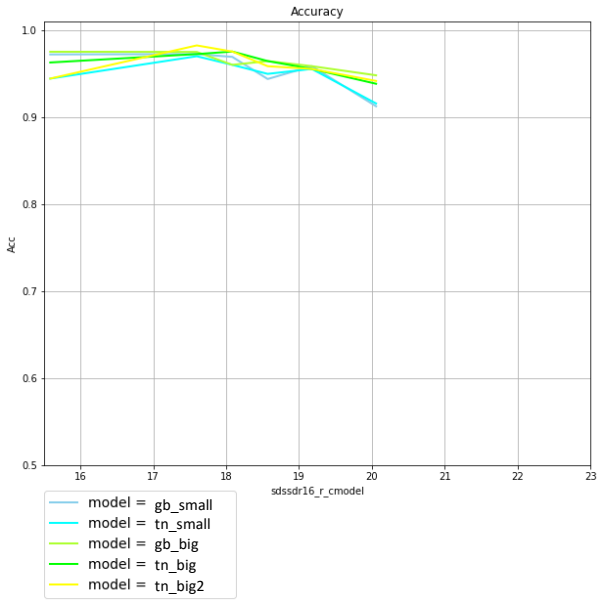


Рисунок 13

Точность моделей на данных (2) в зависимости от яркости.

Таким образом, в рамках данной задачи TabNet не показал превосходства в точности в сравнении с градиентным бустингом.

5.3.2. TabNet лучше бустинга отбирает признаки?

При получении важности признаков, было замечено, что TabNet среди важных также отбирает признаки, не являющиеся физически обоснованными (См. Приложение В)

В дальнейшем это показало, что признаки, отобранные TabNet, не дают точность лучше, выбранных градиентным бустингом:

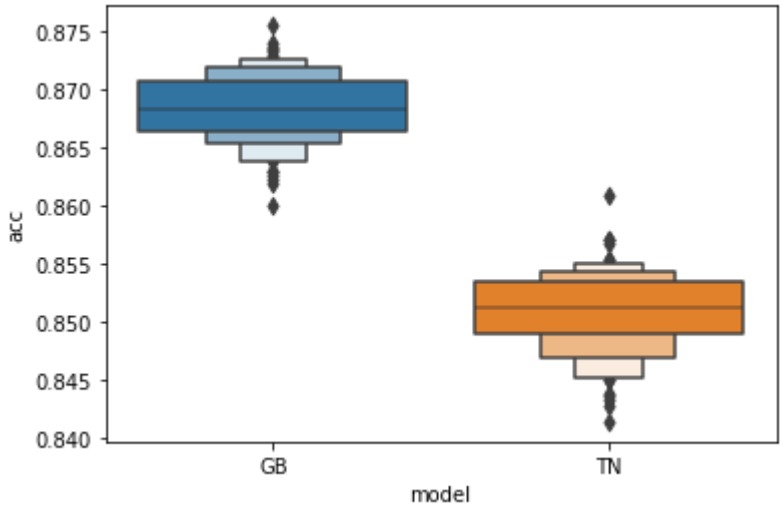


Рисунок 14

Точность моделей градиентного бустинга на бутстрэп выборках из того же набора объектов на 3 самых значимых признака от градиентного бустинга и TabNet соответственно.

Таким образом, TabNet не показывает улучшения при отборе признаков.

5.3.3. TabNet занимает меньше места?

Для проверки размера были построены различные модели TabNet по количеству параметров и модели градиентного бустинга на разных по объему обучающих выборках:

		Размер обучающей выборки			
	Количество параметров	**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		

Затем каждая из них была протестирована на выборке, состоящей из 1912769 объектов, с измерением времени получения предсказания.

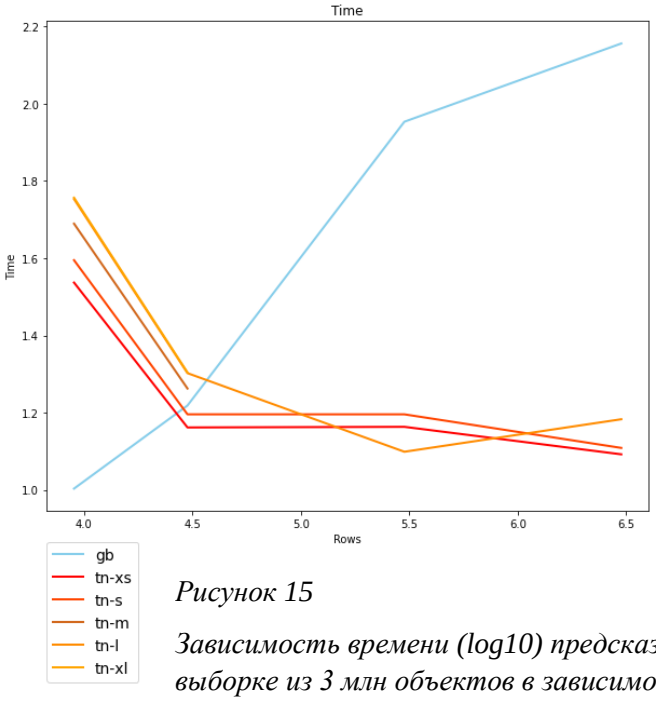


Рисунок 15

Зависимость времени (log10) предсказания на выборке из 3 млн объектов в зависимости от размера обучающей выборки (log10)

Время предсказания в секундах

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		

Была обнаружена существенная разница во времени при увеличении обучающей выборки при сравнимых точностях для моделей TabNet и градиентного бустинга.

За счет независимости количества параметров нейронной сети от величины обучающей выборки, TabNet не увеличивается в размере, в отличие от градиентного бустинга. Таким образом TabNet может быть использован для получения оперативного предсказания при обучении на большом количестве размеченных объектов.

6. Итог

Исследованы и построены ряд моделей для классификации рентгеновских объектов, 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.

Литература

[1] «Measures of Flux and Magnitude SDSS DR12,» [В Интернете]. Available: https://www.sdss.org/dr12/algorithms/magnitudes/.

[2] «The Pan-STARRS1 data archive home page,» [В Интернете]. Available: https://panstarrs.stsci.edu/.

[3] «Data Release Description DESI Legacy Imaging Surveys,» [В Интернете]. Available: https://www.legacysurvey.org/dr8/description/.

[4] «XMMSSC - XMM-Newton Serendipitous Source Catalog (4XMM-DR10),» [В Интернете]. Available: https://heasarc.gsfc.nasa.gov/W3Browse/xmm-newton/xmmssc.html.

[5] C. M. U. Stephanie M. LaMassa, «THE DEG2 RELEASE OF THE STRIPE 82 X-RAY SURVEY: THE POINT SOURCE CATALOG,» [В Интернете]. Available: https://arxiv.org/pdf/1510.00852.pdf.

[6] F. A. A. B. A. Queiroz, «From the bulge to the outer disc: StarHorse stellar parameters, distances, and extinctions for stars in APOGEE DR16 and other spectroscopic surveys,» 2 March 2021. [В Интернете]. Available: https://arxiv.org/pdf/1912.09778v1.pdf.

[7] L. Breiman, «Random Forests. Machine Learning,» *Machine Learning,* № 45, p. 5–32, 2001.

[8] A. K. Alexey Natekin, «Gradient boosting machines, a tutorial,» *Front Neurorobot,* № 4, pp. 7-21, 2013.

[9] «LightGBM’s documentation,» [В Интернете]. Available: https://lightgbm.readthedocs.io/en/latest/index.html.

[10] T. P. Sercan O. Arık, «TabNet: Attentive Interpretable Tabular Learning,» [В Интернете]. Available: https://arxiv.org/pdf/1908.07442.pdf.

[11] E. F. S. a. D. P. Finkbeiner, «MEASURING REDDENING WITH SLOAN DIGITAL SKY SURVEY,» *The Astrophysical Journal STELLAR SPECTRA AND RECALIBRATING SFD,* № 13pp, pp. 737-103, 2011.

[12] D. J. S. Arjun Dey, «Overview of the DESI Legacy Imaging Surveys,» *The Astronomical Journal,* № 29pp, pp. 157-168, 2019.

[13] Wai, «An Example of Hyperparameter Optimization on XGBoost, LightGBM and CatBoost using Hyperopt,» 2 Aug 2019. [В Интернете]. Available: https://towardsdatascience.com/an-example-of-hyperparameter-optimization-on-xgboost-lightgbm-and-catboost-using-hyperopt-12bc41a271e.

[14] A. M. M. S. R. G. a. V. G. A. O. Clarke, «Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra,» *A&A,* т. 639, № A84, p. 29, 2020.

Приложение А

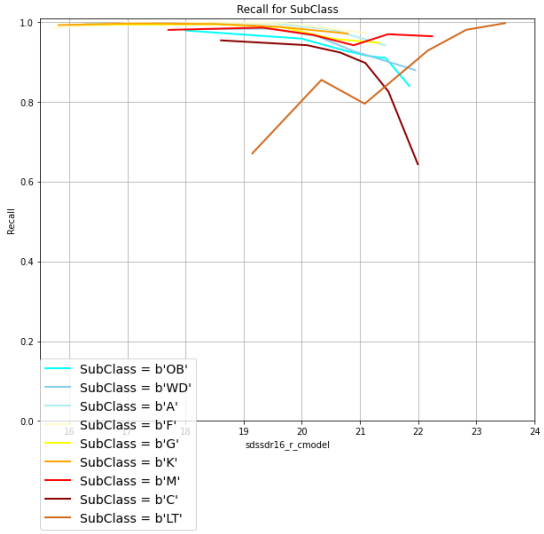
Гиперпараметры лучших моделей:

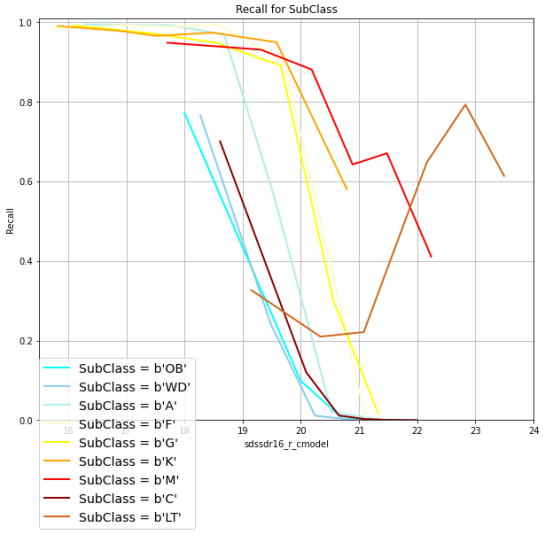
	gb_small	gb_big
min_child_samples	49	31
colsample_bytree	0.7287958151148961	0.7612466445456562
num_leaves	93	35
min_child_weight	0.471460620399453	0.2820906688539359

	Число параметров	**Nd=Na**	**λsparse**	**γ**	**Nsteps**	**shared**	**decision**
tn_small	9296	8	0.003	1.5797	4	1	2
tn_big	10128	8	0.008	2.3697	4	2	2
tn_big2	158296	32	0.008	2.3696	4	3	3
tn_big_and_small	158296	32	0.008	2.3697	4	3	3

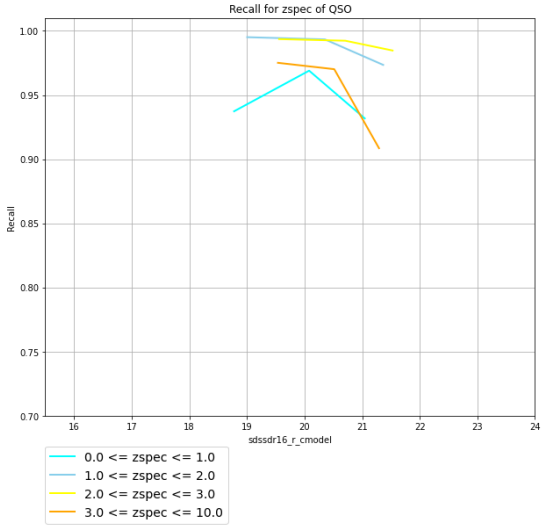
Приложение Б

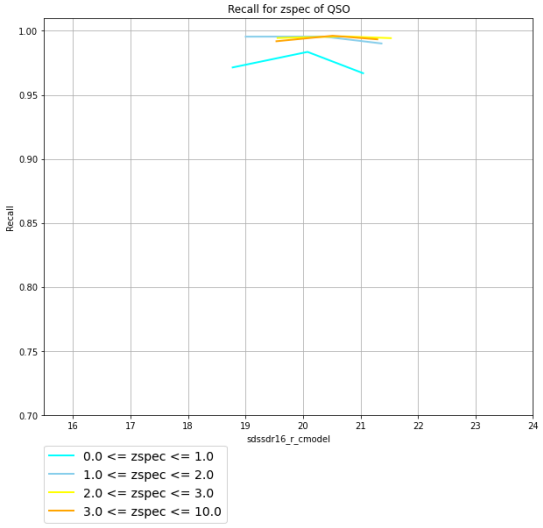
Recall для подклассов в зависимости от признака.
Для звезд:



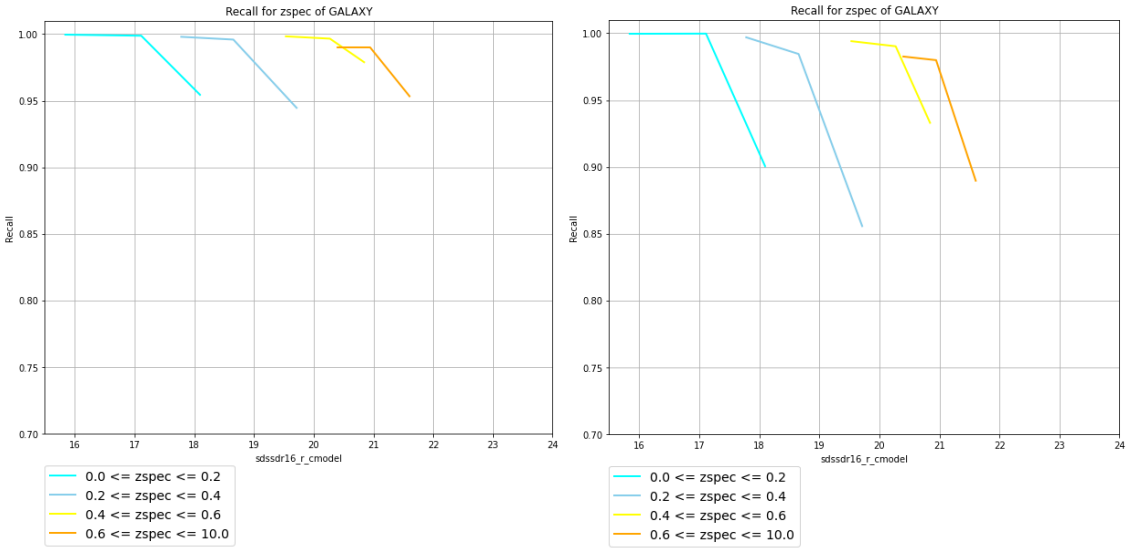


Для квазаров:





Для галактик:



Для всех графиков:

- (a) – для моделей, обученных на большой выборке,
- (b) – для моделей, обученных на 15-ти тыс рентгеновских объектов.

Приложение В

Отбор важных признаков TabNet и градиентным бустингом: