Тема: "Исследование и разработка алгоритмов машинного обучения для построения трехмерной карты рентгеновских звезд в Галактике Млечный Путь"

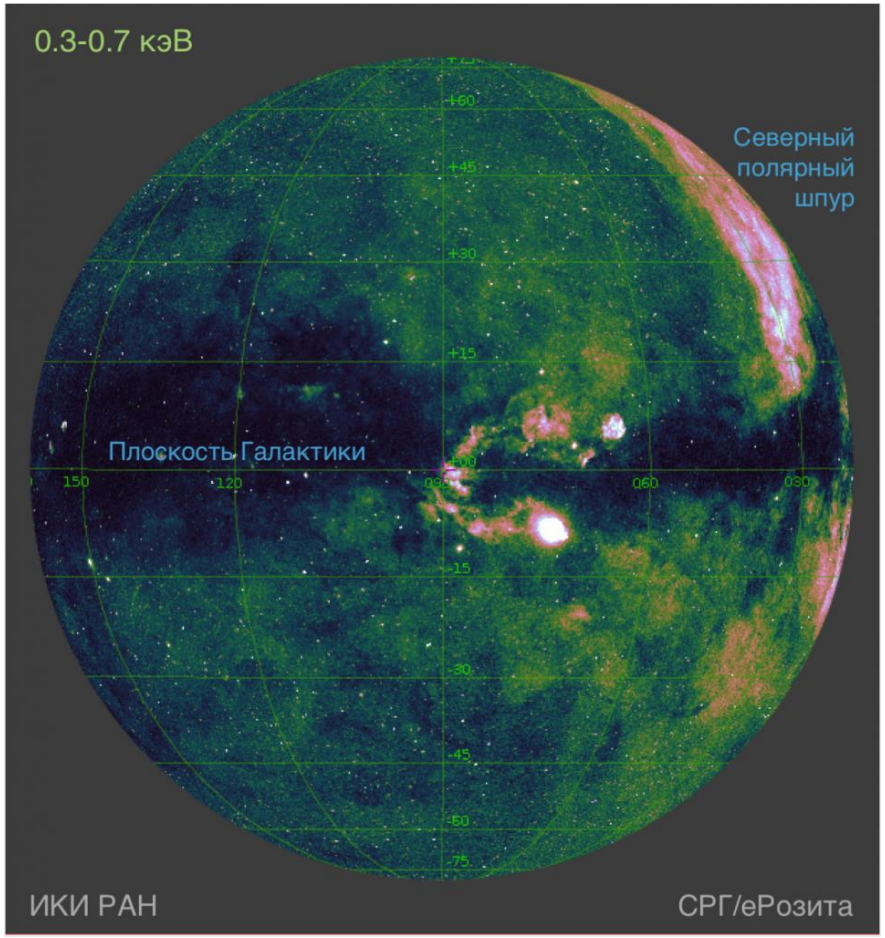
Малышева Надежда
320 группа

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

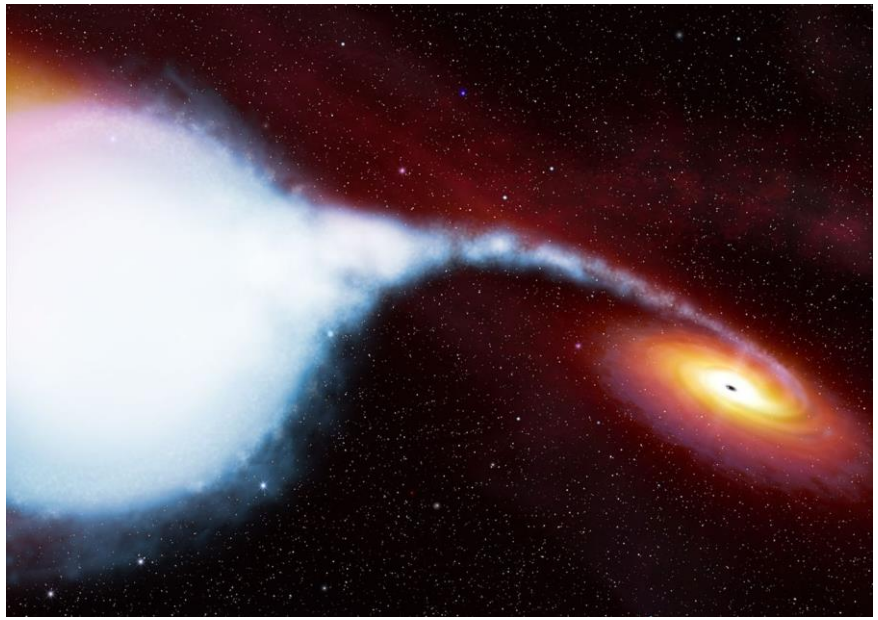
К вечеру 11 июня телескоп СРГ/eРозита завершил построение карты, охватывающей всю небесную сферу, площадь которой составляет 41 тысячу 253 квадратных градуса.

На этих картах зарегистрировано около полумиллиона рентгеновских источников, в том числе звезды.

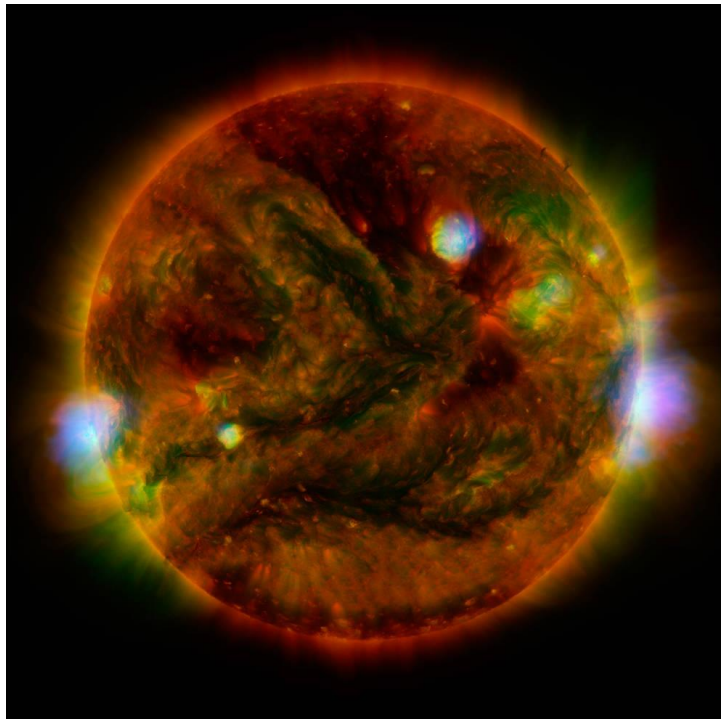
Введение:

Рентгеновскими звездами принято считать объекты имеющую большую светимость в рентгеновском диапазоне.

Рентгеновские двойные звезды.
Нас интересует случай, когда пара представлена в виде массивной звезды и компактным объектом.



Обычные звезды с рентгеновскими вспышками



Актуальность:

Рентгеновскими звездами принято считать объекты имеющую большую светимость в рентгеновском диапазоне.

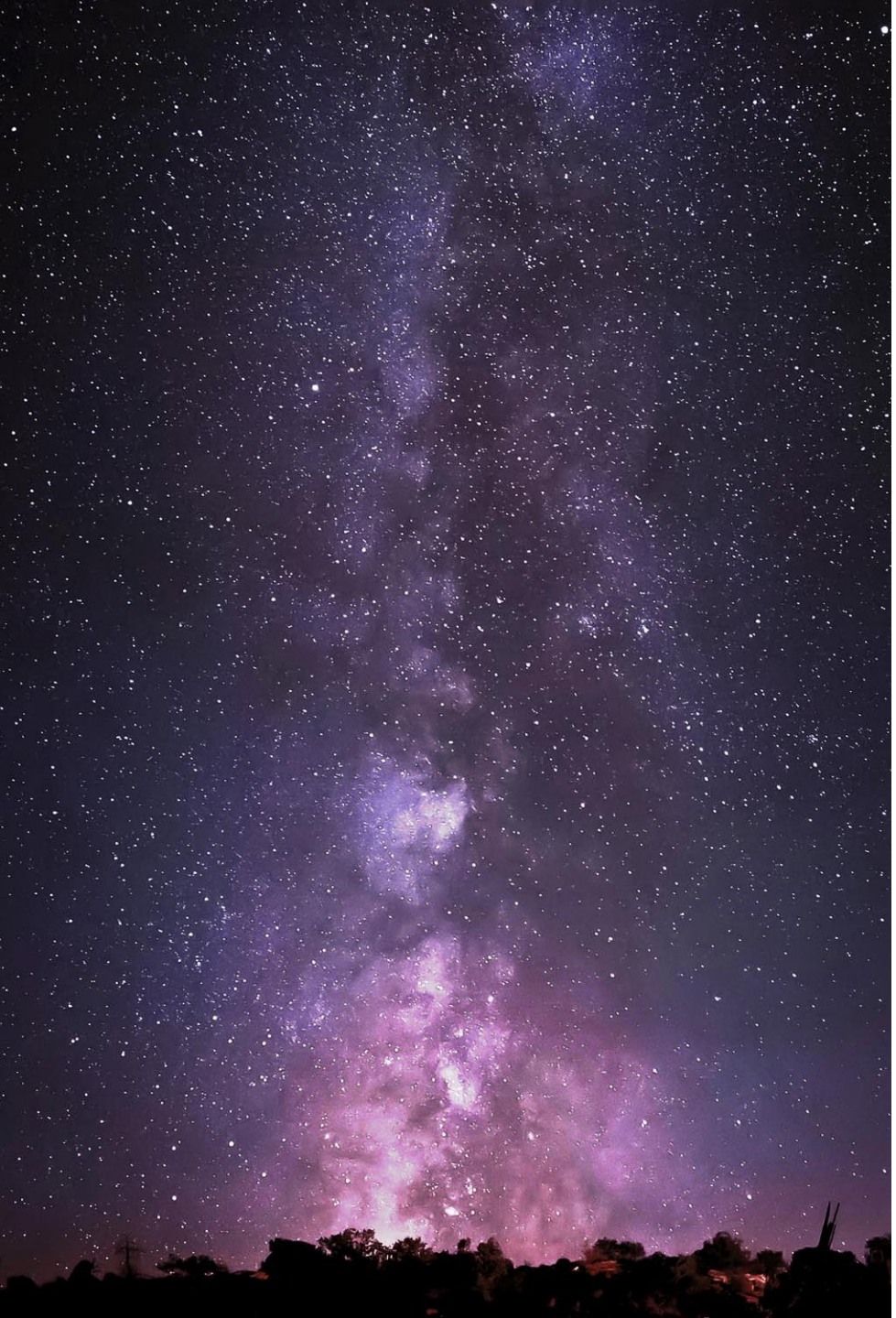
Рентгеновские двойные звезды.
Нас интересует случай, когда пара представлена в виде массивной звезды и компактным объектом.

Интересны процессы происходящие радом с компактными объектами. Таким образом можно исследовать физику экстремального состояния, которую невозможно воссоздать в лабораторных условиях.

Поэтому важно их обнаруживать и классифицировать такие объекты.

Обычные звезды

Исследование активности обычных звезд в рентгеновском диапазоне. Это важно для понимания, около каких звезд возможна жизнь: наличие большой рентгеновской активности говорит, о невозможности существования жизни рядом с ними.



Актуальность задачи :

Из-за того, что мы находимся на диске, а не смотрим со стороны, структуру нашей Галактики трудно изучать. На данный момент до сих пор нет точных данных о спиральных рукавах.

Зачем это знать?

- Это позволит нам понять, насколько другие галактики похожи на нашу, насколько мы можем обобщать имеющиеся данные о них, на Млечный Путь, на нашу звездную систему.

- А также в спиральных рукавах образуются звезды. Там повышенный темп звездообразования, поэтому все, что связано с молодыми звездами — должно концентрироваться в спиральных рукавах, что так же интересно для учёных.

Актуальность задачи :

Рентгеновские звезды концентрируются к спиральным рукавам. Зная данные о расположении рентгеновских звезд, мы сможем получить данные о спиральных рукавах нашей галактики.

Рентгеновские звезды являются очень мощными, заметными источниками, которые мы можем видеть практически с другой стороны галактики, но сложно точно определить расстояние до них.

А эти данные мы можем узнать только с помощью методов прогнозирования

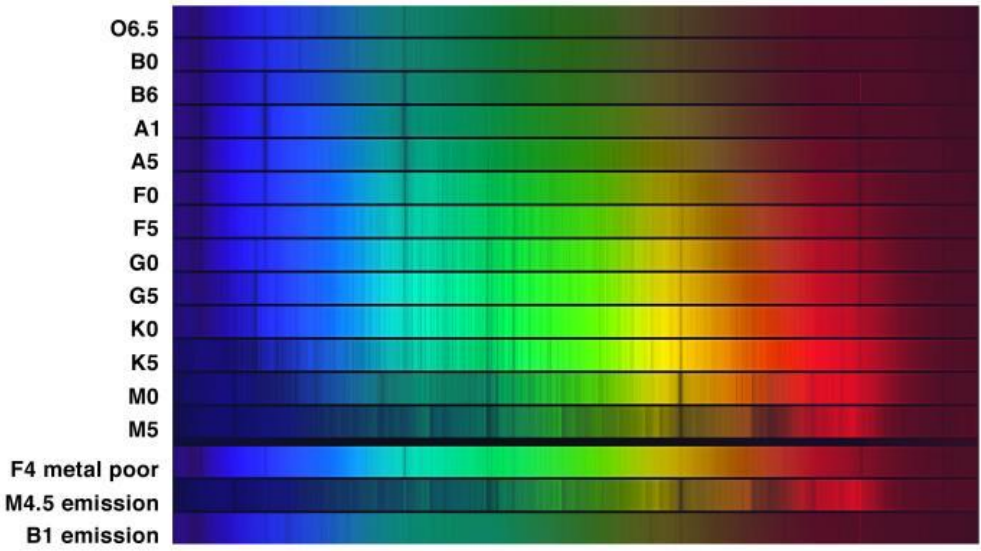
Актуальность:

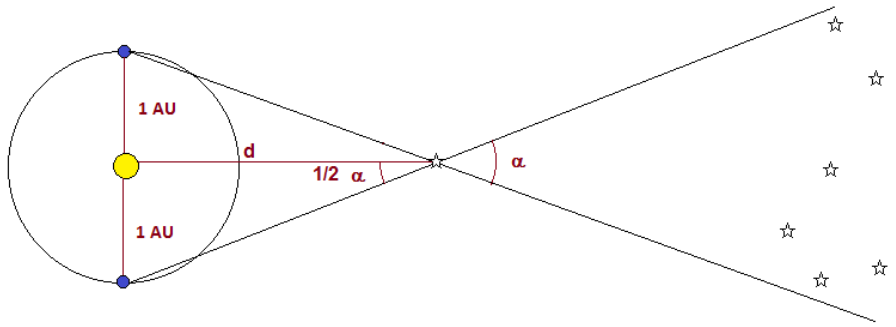
Почему машинное обучение?

Существует ограниченное число способов классификации объектов и измерения расстояния до него.

- Класс звезды можно получить по спектру, но это трудно сделать для слабых объектов, которые возможно только сфотографировать. Поэтому необходимо развивать фотометрические методы оценки класса.

- Расстояние до объекта можно изучать с помощью параллакса. Но это возможно для далеких объектов





Постановка задачи:

Исследовать и разработать модели машинного обучения для классификации рентгеновских звезд и измерение расстояния до них

Задача разбивается на несколько подзадач:

1)Построение классификатора звезд

2)Построение классификатора конкретного спектрального класса

- Против Галактик и Квазаров
- Против других звезд

3)*Построение модели регрессии для предсказания параллакса звезд

Обзорная часть:

Методы ML:

- Логистическая регрессия
- K-ближайших соседей (KNN)
- Метод опорных векторов (SVM)
- Случайный лес
- Градиентный бустинг

Случайный лес и градиентный бустинг дают наиболее точные результаты, поэтому рассмотрим их подробнее:

Изученные статьи:

- Leo Breiman «Random Forests»
- Lin & Jeon «Random Forests and Adaptive Nearest Neighbors»
- Natekin (2013) «Gradient boosting machines, a tutorial»
- «Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms»

Случайный лес:

Случайный лес — это классификатор, состоящий из классификаторов 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:

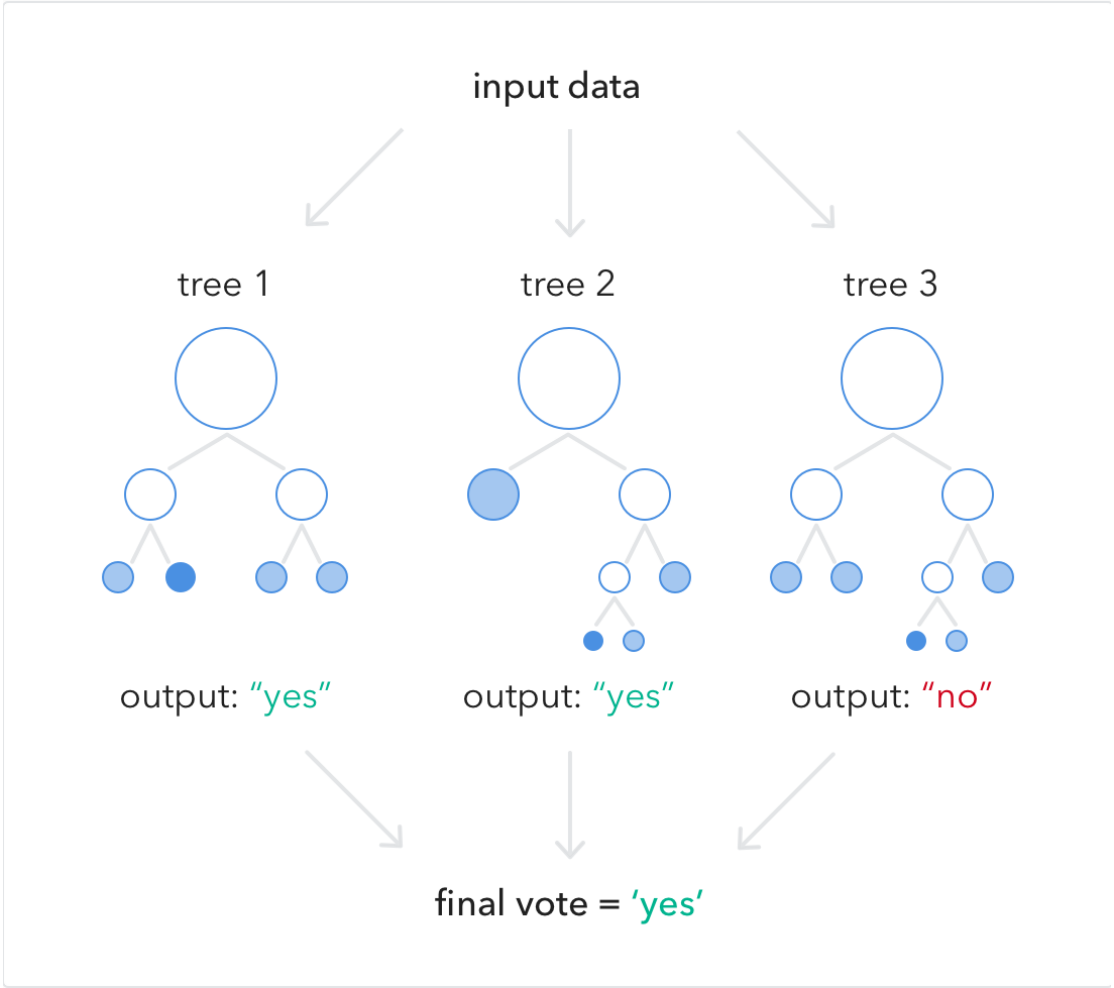
- Сформировать бутстрэп выборку S, размера F по исходной обучающей выборке
- По выбранной S выборке строится дерево решений без ограничения глубины с расщеплением каждой вершины дерева только по фиксированной доле случайно отбираемых признаков.

2) В результате получаем ансамбль из k деревьев решений

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

Достоинства:

- Хорошая точность (как у Adaboost) (т. к. деревья в ансамбле слабо коррелируемы)
- Устойчив к выбросу и шуму
- Быстрый (за счет обучения каждого дерева на части данных)
- Легкость организации параллельных вычислений
- Не переобучается



Недостатки:

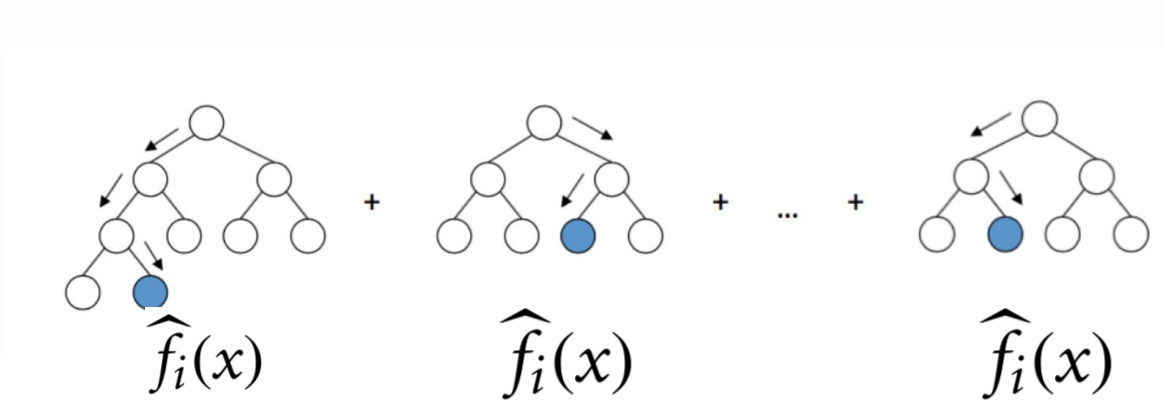
- Не интерпретируемые модели
- Плохо работает на разреженных признаках
- Большой размер

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
- Простота

Недостатки:

- Трудоемкий метод, занимает много памяти
- 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

Обзорная часть:

Данные:

SDSS обзор - проект широкомасштабного исследования многоспектральных изображений и спектров красного смещения звёзд и галактик при помощи 2,5-метрового широкоугольного телескопа.

Изображения снимались с помощью фотометрической системы из пяти фильтров, которые имеют названия **u**, **g**, **r**, **i** и **z**.

Используемая выборка SDSS фотометрических данных состоит из 4614588 наблюдений с 10-ю фотометрическими признаков.
13731 OB звезд
946632 других звезд
2789052 Галактик
86517 Квазаров

Построение решения:

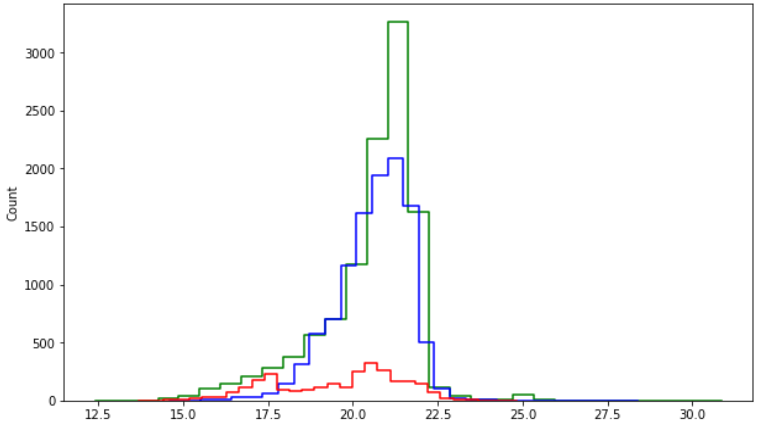
Решим более локальную задачу для самых мощных звезд:

- Классификация OB против галактик и квазаров
- Классификация OB против других звезд

Построение решения:

Классификатор OB звезд от Галактик и Квазаров:

OB vs G, Q	Объекты 20% - test	Распределение OB:G:Q
Train	24714	1 : 0,25 : 1
test	5493	1 : 0,5 : 0,5



- В обучающей выборке добавлено больше квазаров, т. к. их труднее всего отделить
- Нормализация данных

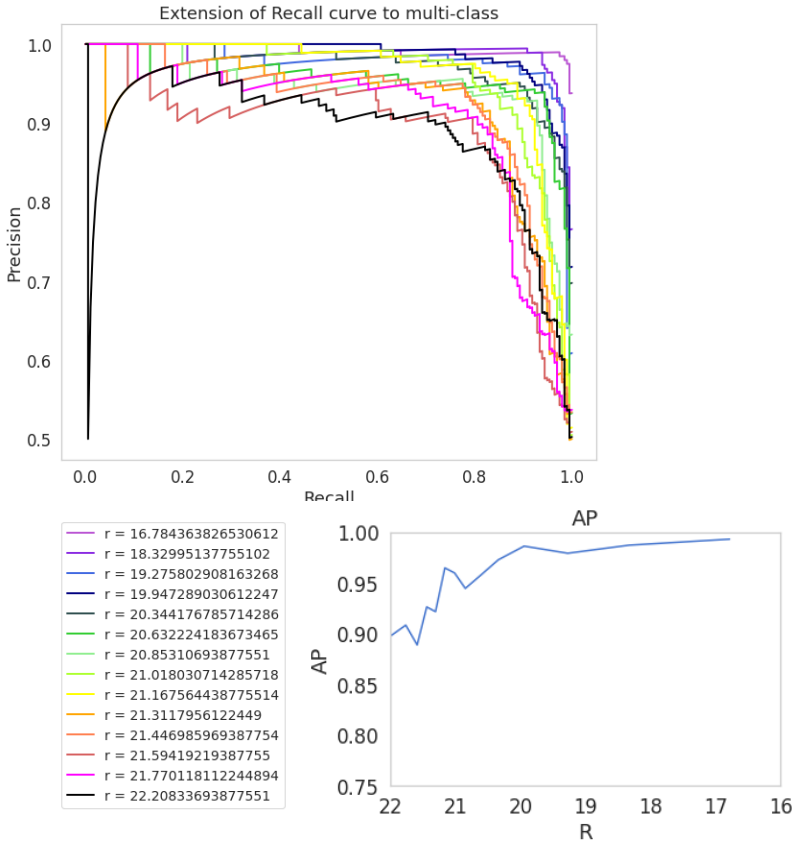
	Iterations	Depth	Learn. Rate	Features	Samples leaf
Градиентный бустинг	150- 3000	1 - 16	0,001 – 0,1	-	5 - 60
Случайный лес	50 - 300	1 - 20	-	1 - 10	1 - 14

Построение решения:

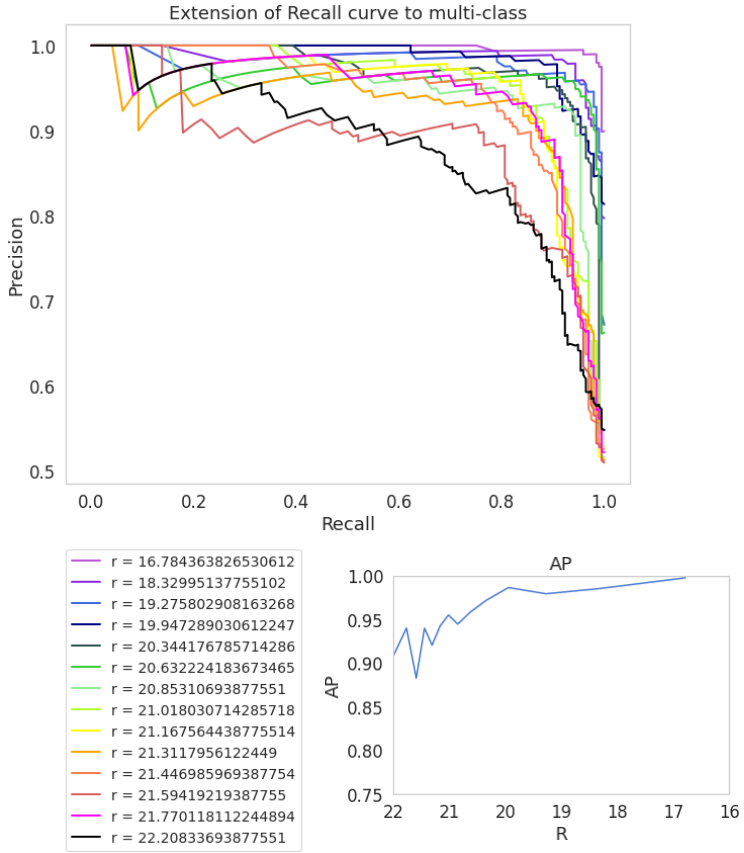
Классификатор ОВ звезд от Галактик и Квазаров:

Результаты:

Градиентный бустинг



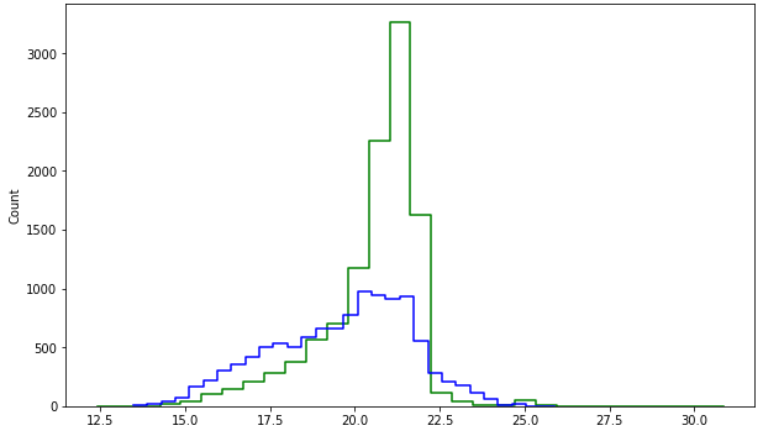
Случайный лес



Построение решения:

Классификатор OB звезд от звезд других спектральных классов:

OB vs STAR	Объекты 20% - test	Распределение OB : другие звезды
Train	21960	1:1 (равномерно для каждого подкласса)
test	5499	1:1

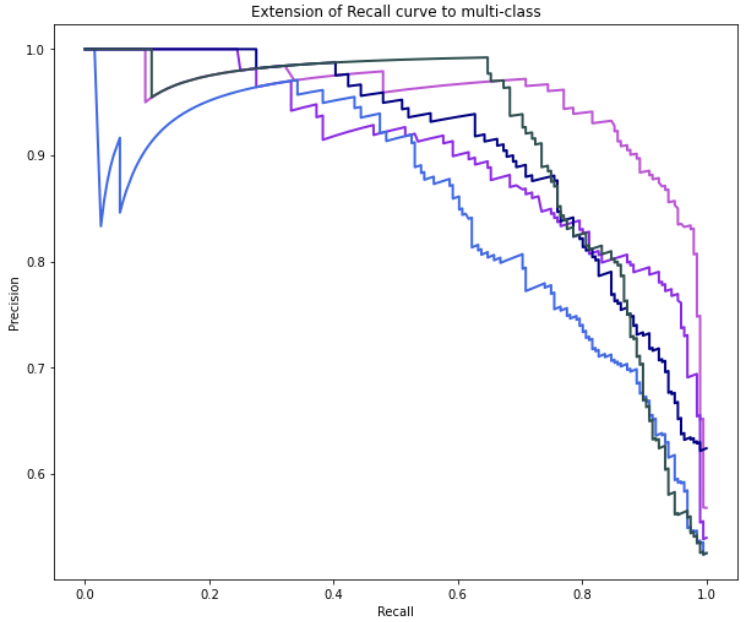


- Объединение основных подклассов
- В тестовой и обучающей выборке равномерно распределены подклассы других звезд по количеству
- Нормализация данных

Построение решения:

Классификатор OB звезд от звезд других спектральных классов:

Градиентный бустинг



r = 16.784363826530612
r = 18.32995137755102
r = 19.275802908163268
r = 19.947289030612247
r = 22.20833693877551

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.88	0.84	2747
1	0.87	0.79	0.83	2752
accuracy			0.84	5499
macro avg	0.84	0.84	0.84	5499
weighted avg	0.84	0.84	0.84	5499

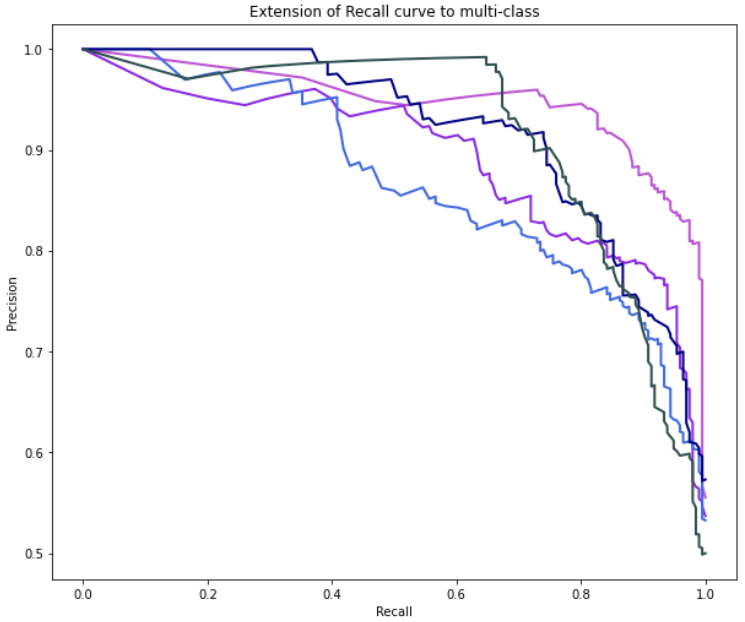
Confusion Matrix:

```
 [[2427  320]
 [ 578 2174]]
```

Training Score: 0.8664845173041894

Testing Score: 0.8366975813784324

Случайный лес



r = 16.784363826530612
r = 18.32995137755102
r = 19.275802908163268
r = 19.947289030612247
r = 22.20833693877551

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.87	0.84	2747
1	0.86	0.79	0.82	2752
accuracy			0.83	5499
macro avg	0.83	0.83	0.83	5499
weighted avg	0.83	0.83	0.83	5499

Confusion Matrix:

```
 [[2389  358]
 [ 580 2172]]
```

Training Score: 1.0

Testing Score: 0.8294235315511911

26.750158548355103

Планы:

- Улучшение существующих классификаторов
- Добавление данных
- Построение модели регрессии для предсказания параллакса звезд
- Изучение влияния поглощения на предсказание

Спасибо за внимание!