|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

**Факультет «Информатика и системы управления»**

**Кафедра «Системы обработки информации и управления»**

Отчет по рубежному контролю №2

**«Методы построения моделей машинного обучения»**

по дисциплине «Технологии машинного обучения»

Вариант №27

Выполнил:

студент группы ИУ5Ц-84Б   
Папин А.В.

подпись, дата

Проверил:

к.т.н., доц., Ю.Е. Гапанюк

подпись, дата

2024 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Задание 3](#_Toc17)

[1. Листинг 4](#_Toc18)

[1.1. Подключение библиотеки и получение датасета 4](#_Toc19)

[1.2. Изучение данных 5](#_Toc20)

[1.4. Пропущенные значения 8](#_Toc21)

[1.5. Дублирующие значения 8](#_Toc22)

[1.6. Устранение сильных выбросов 8](#_Toc23)

[1.7. Машинное обучение 9](#_Toc24)

[1.8. Деление на обучающий и тестовой выборки 9](#_Toc25)

[1.9. Обучение модели 11](#_Toc26)

[1.9.1. LogisticRegression 12](#_Toc27)

[1.9.2. LGBMClassifier 16](#_Toc28)

[1.9.3. CatBoostClassifier 20](#_Toc29)

[1.9.4. XGBClassifier 23](#_Toc30)

[1.9.5. GradientBoostingClassifier 27](#_Toc31)

[1.10. Анализ моделей 30](#_Toc32)

[1.11. Вывод 31](#_Toc33)

[1.12. Гистограмма 32](#_Toc34)

## **Задание**

Для заданного набора данных (по Вашему варианту) постройте модели классификации или регрессии (в зависимости от конкретной задачи, рассматриваемой в наборе данных). Для построения моделей используйте методы 1 и 2 (по варианту для Вашей группы). Оцените качество моделей на основе подходящих метрик качества (не менее двух метрик). Какие метрики качества Вы использовали и почему? Какие выводы Вы можете сделать о качестве построенных моделей? Для построения моделей необходимо выполнить требуемую предобработку данных: заполнение пропусков, кодирование категориальных признаков, и т.д.

При решении задач можно выбирать любое подмножество признаков из приведенного набора данных.

Для сокращения времени построения моделей можно использовать фрагмент набора данных (например, первые 200-500 строк).

Методы 1 и 2 для каждой группы приведены в следующей таблице:

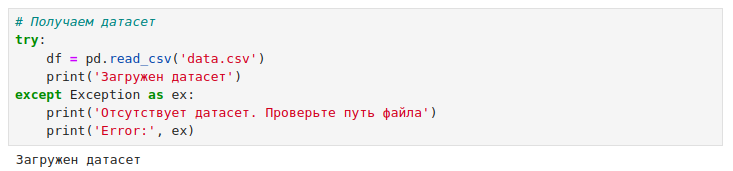
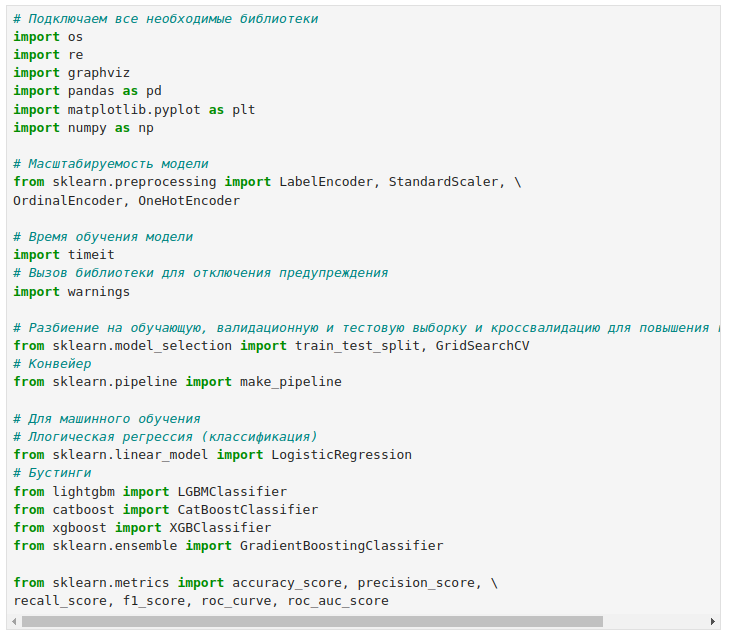
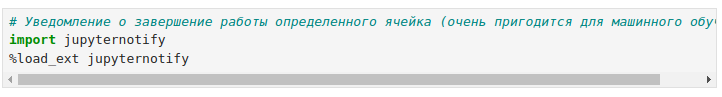
|  |  |  |
| --- | --- | --- |
| **Группа** | **Метод №1** | **Метод №2** |
| ИУ5Ц-84Б | Линейная/логистическая регрессия | Градиентный бустинг |

**Наборы данных:**

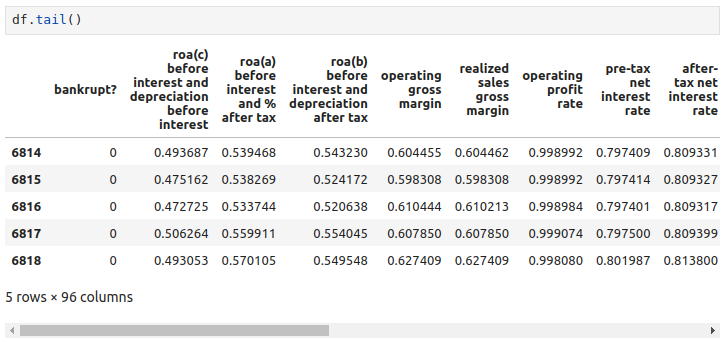
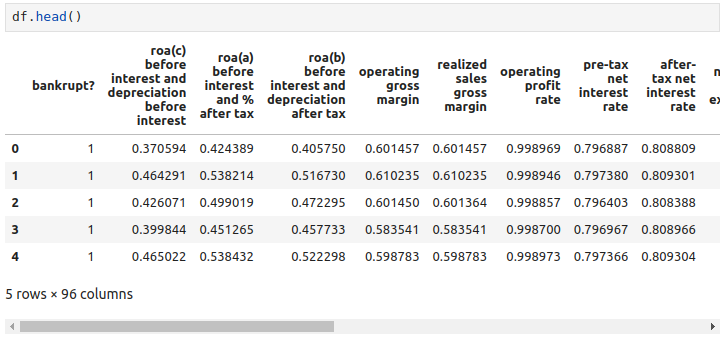
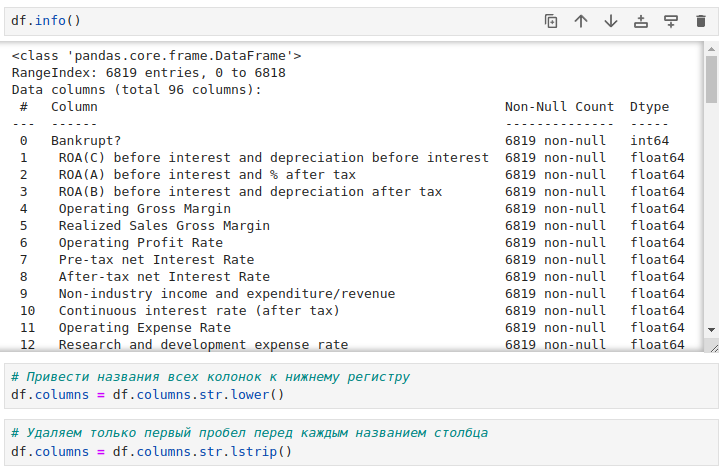
https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction/data

# **Листинг**

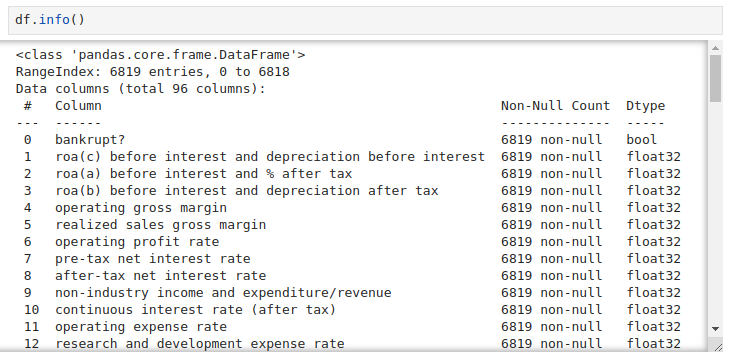
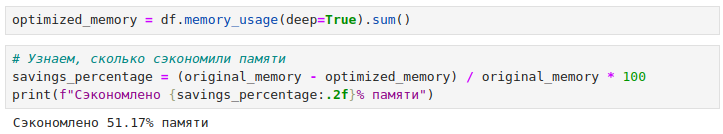
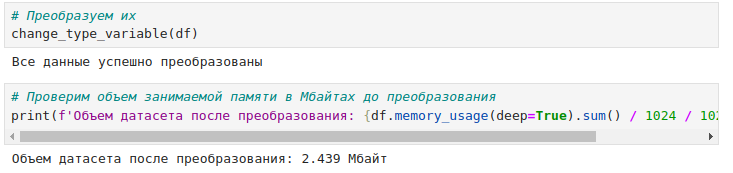
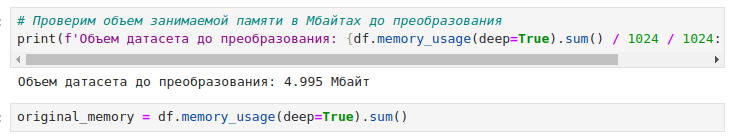
# **Подключение библиотеки и получение датасета**



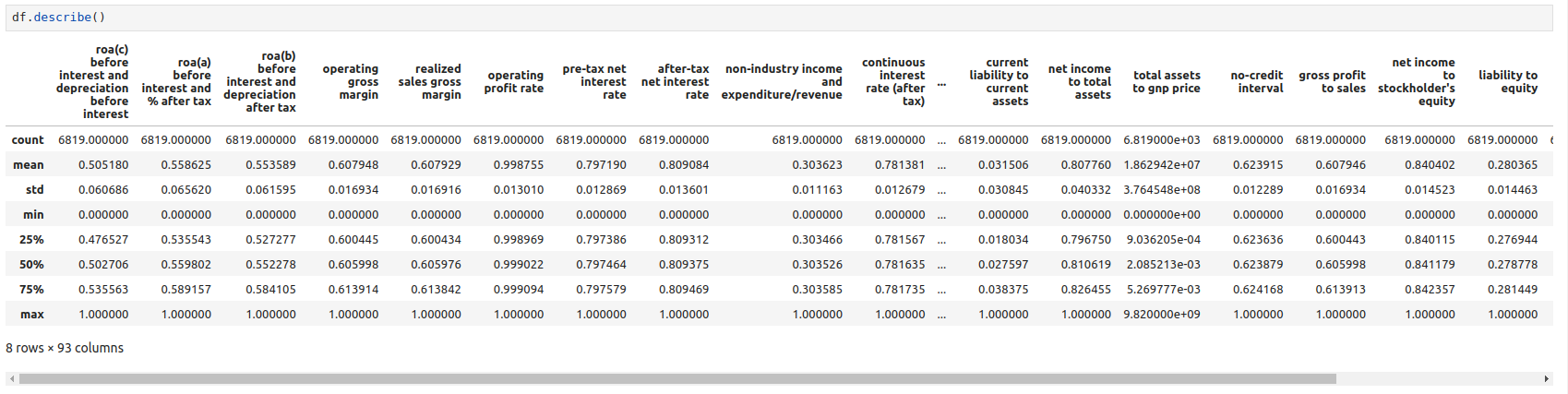
# **Изучение данных**



* 1. **Преобразование данных**



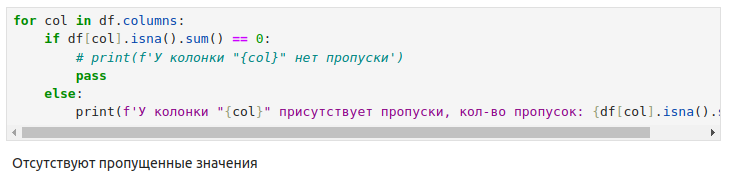
Рассмотрим описательную статистику



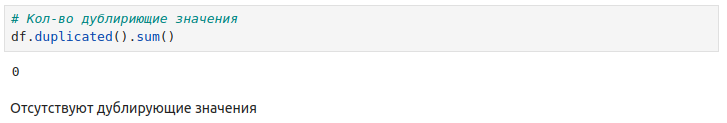
Анализ описательной статистики

* В выборке представлены данные о 6819 компаниях.
* Среди этих компаний 2201 обанкротились (т.е., метка "bankrupt?" равна 1), а 4618 не обанкротились (метка "bankrupt?" равна 0).
* Различные показатели отдачи от активов (ROA) варьируются от 0.37 до 0.57, т.е. операционная прибыль, отношение долга к активам и другие, имеют широкий диапазон значений.
* Коэффициенты операционной прибыли, пред-налоговой чистой процентной ставки и пост-налоговой чистой процентной ставки также имеют широкий диапазон значений.
* Важные финансовые показатели, такие как отношение долга к активам, отношение собственного капитала к обязательствам и степень финансового рычага, также имеют различные значения в выборке.
* Все показатели представлены в нормализованных значениях от 0 до 1, что облегчает их сравнение и анализ.

# **Пропущенные значения**

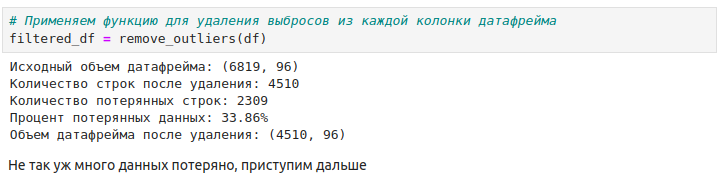


# **Дублирующие значения**

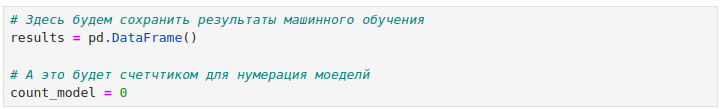


# **Устранение сильных выбросов**

Полное устранние выбросов потребует внимательную и непростую работу, т.к. каждая колонка имеет индивидуальные значения и нужно найти особый подход к ним, поэтому будем устранять только те сильные выбросы, стараясь минимизировать количество выбросов

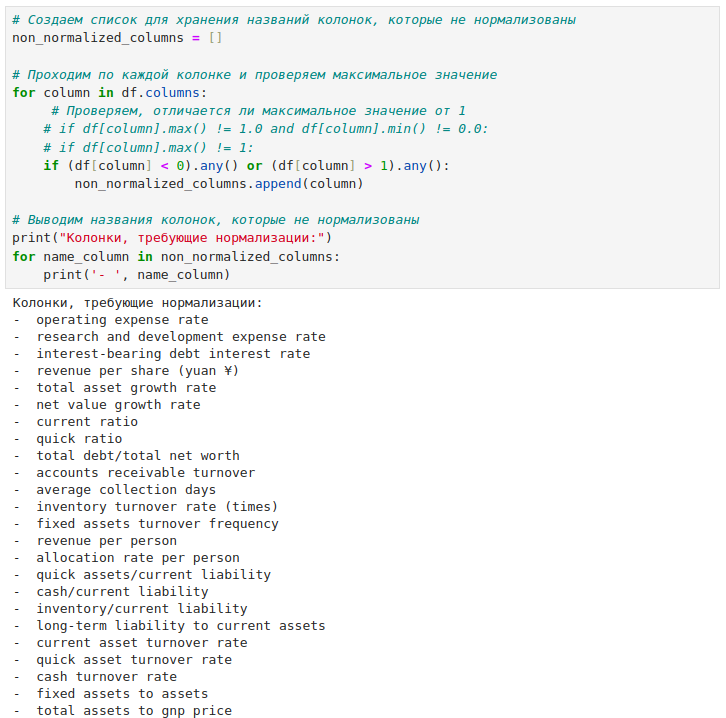


# Машинное обучение

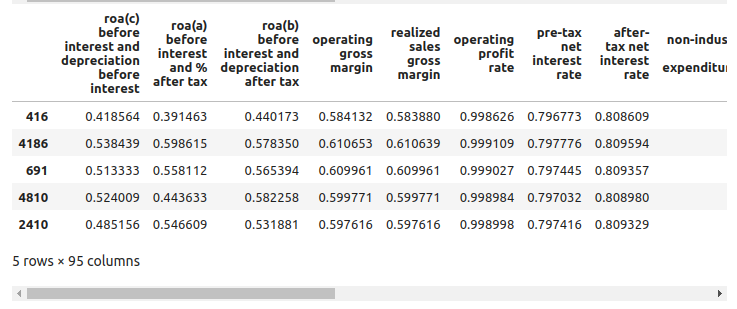
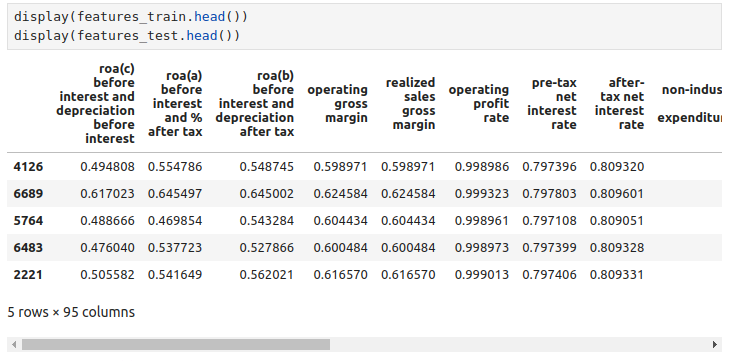
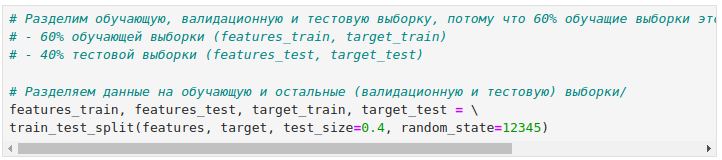
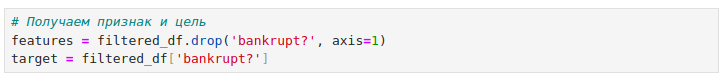


# Деление на обучающий и тестовой выборки

Поскольку у нас данные не нормализованы, т.е. на первый взгляд кажется, что значения находится в диапазонах от 0 до 1, и нет необходимости масштабировать признаков, однако существуют колонки, которые не нормализованы



Как и видим, что есть колонки, поэтому будем масштабировать



# Обучение модели

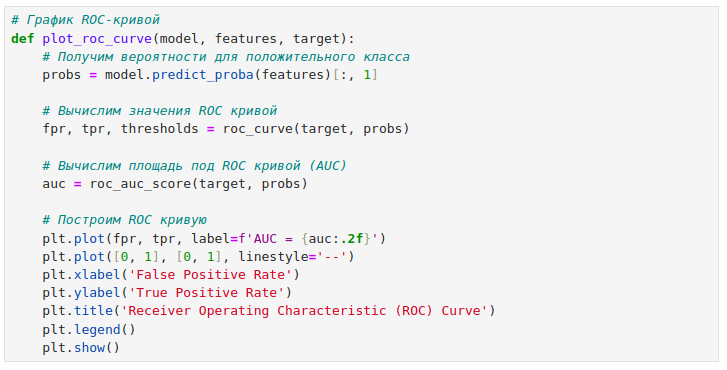
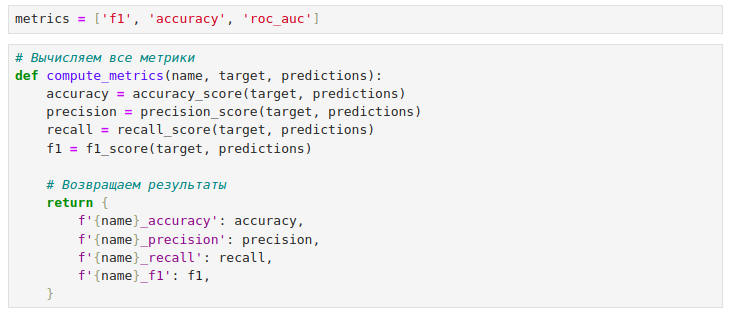
Выберем метрику F1-score, Accuracy и ROC-curve

Почему выбрал эти?

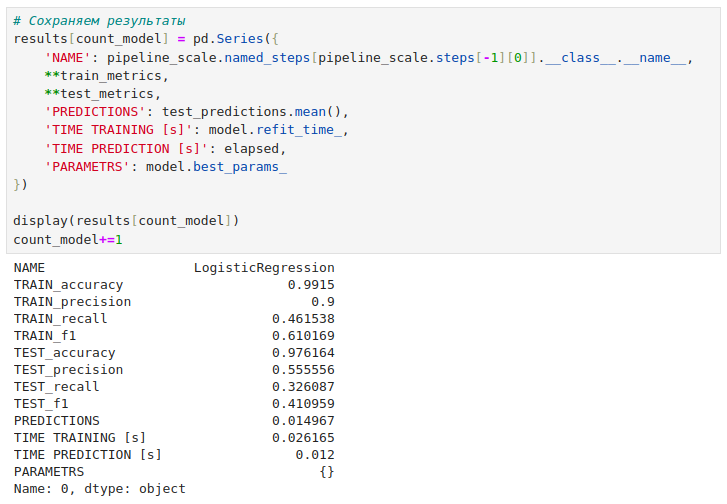
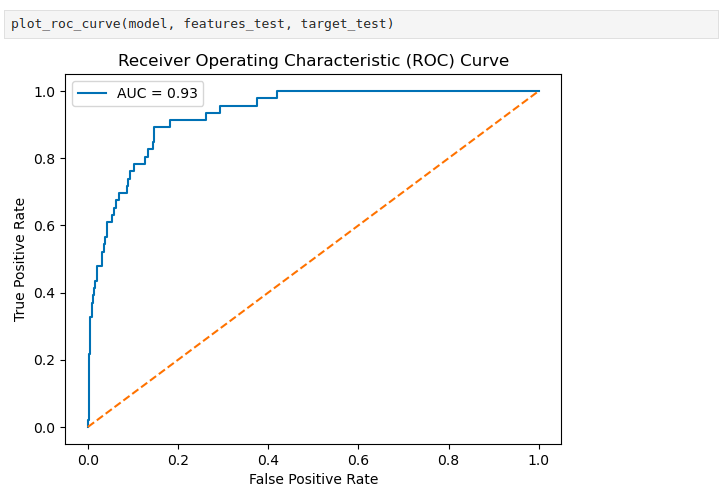
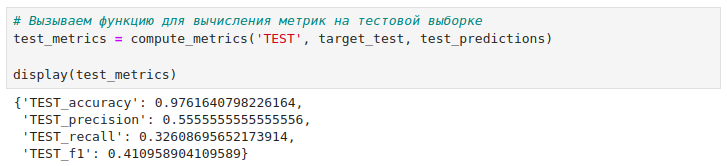
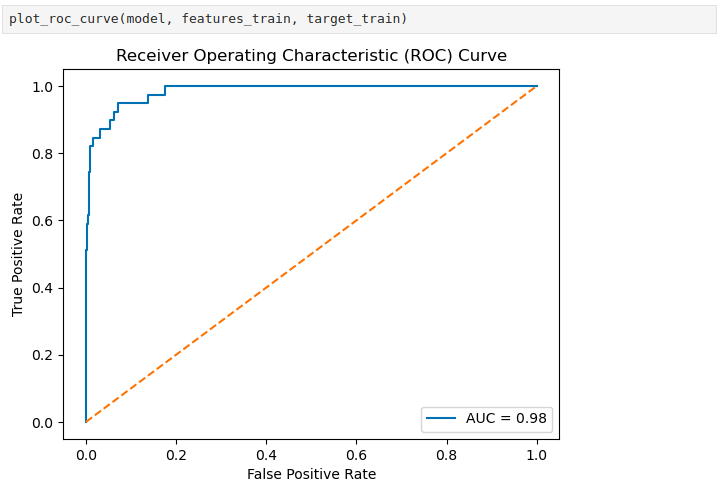
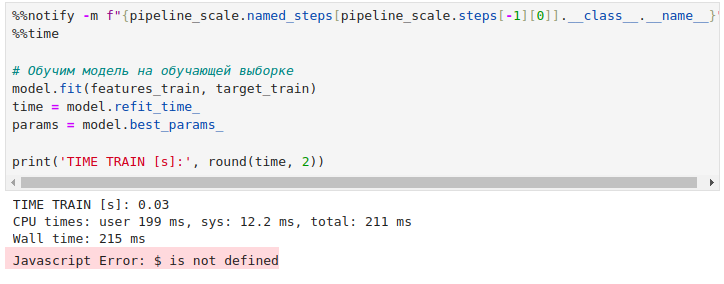
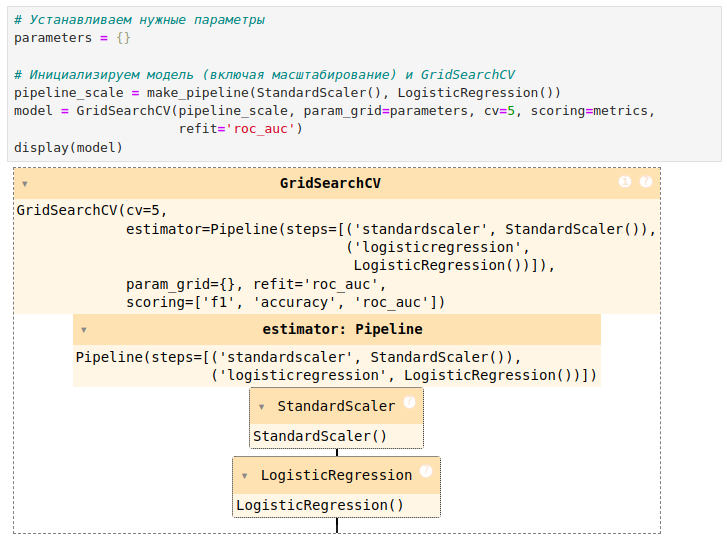
F1-score это хорошая метрикая для баланса между точностью и полнотой в задачах классификации, особенно в случаях, когда классы несбалансированы

Accuracy это, очевидно, простая и понятная метрика, которая показывает общую долю правильно классифицированных экземпляров в общем количестве экземпляров

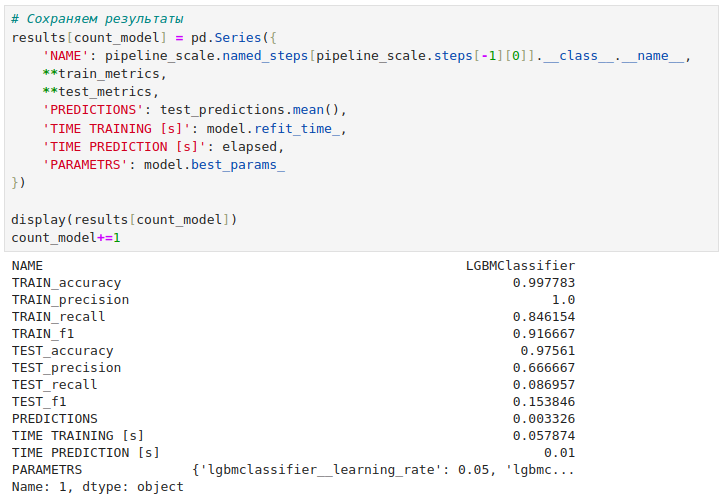
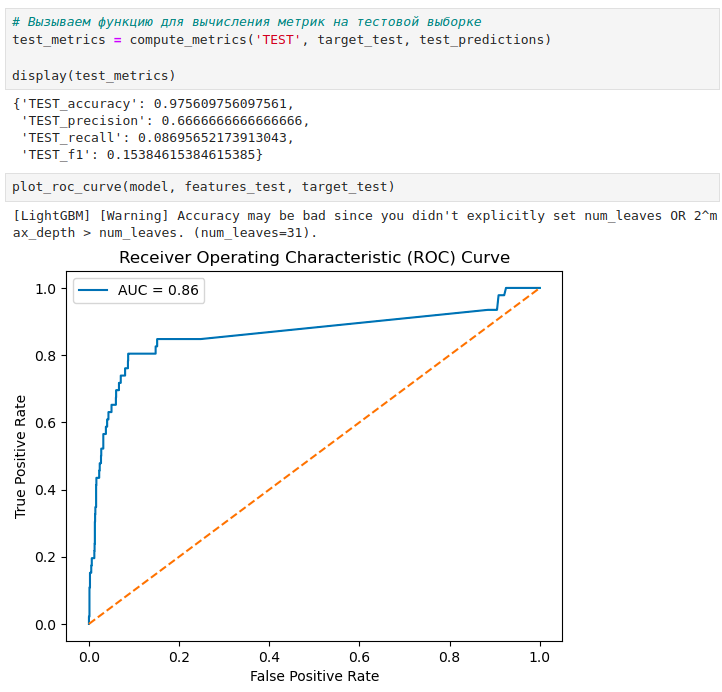
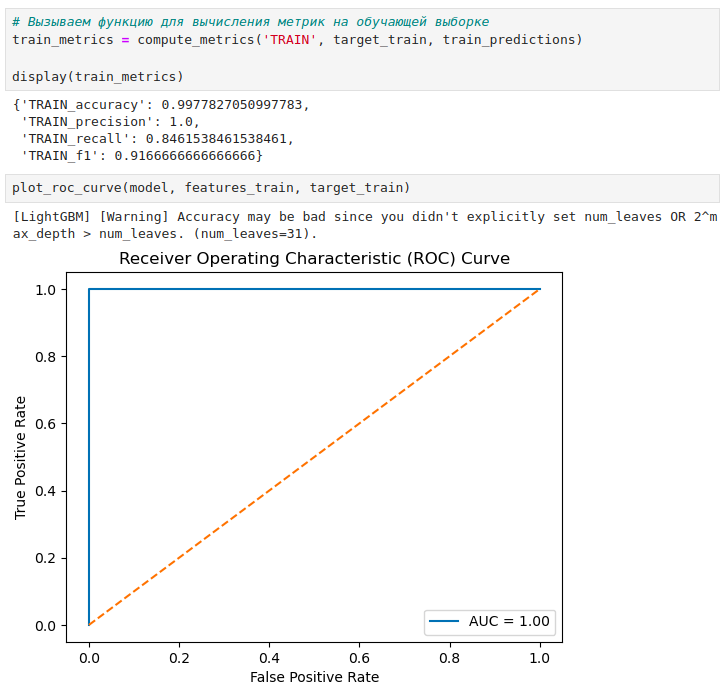
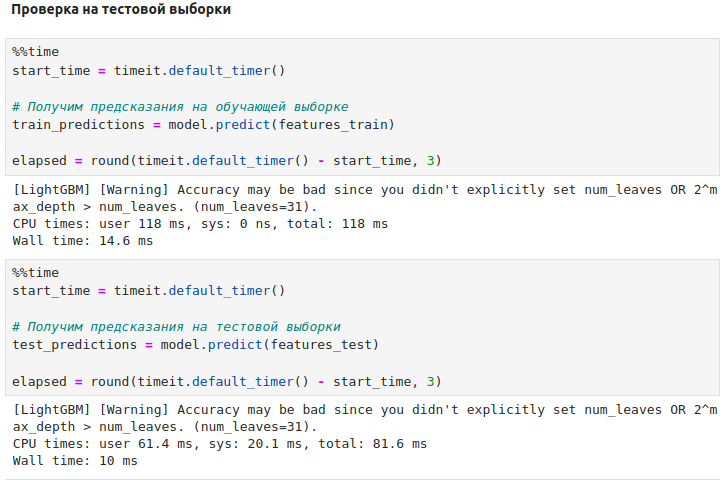
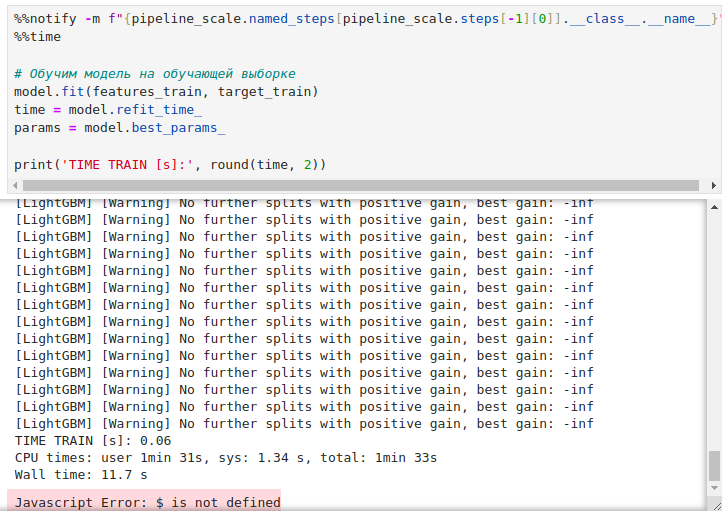
ROC AUC это метрика качества классификатора, которая учитывает все пороговые значения и позволяет оценить способность модели отличать между положительными и отрицательными классами. Например, будет полезно в тех случаях, когда данные несбалансированы или когда интересует только качество ранжирования классификатора



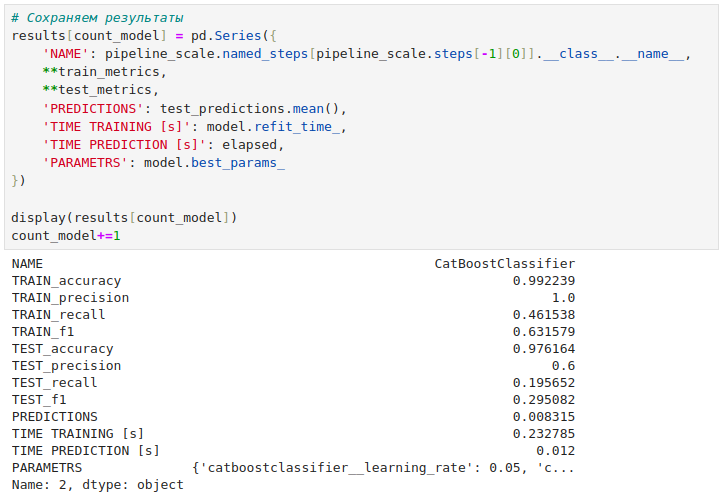
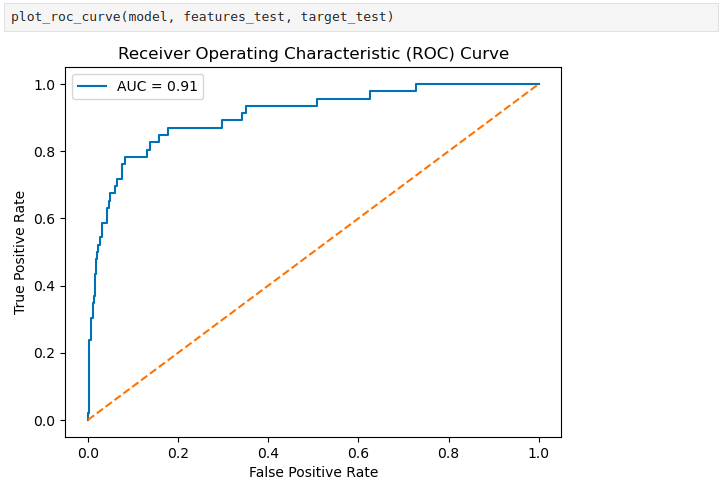
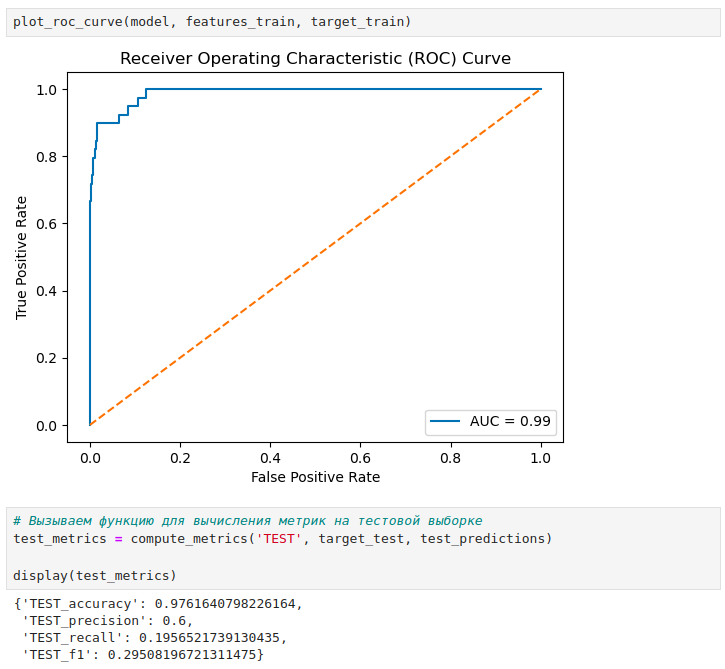
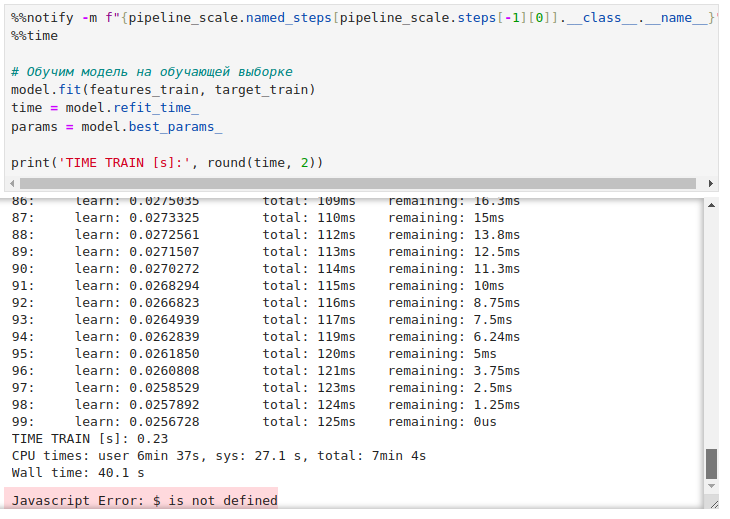
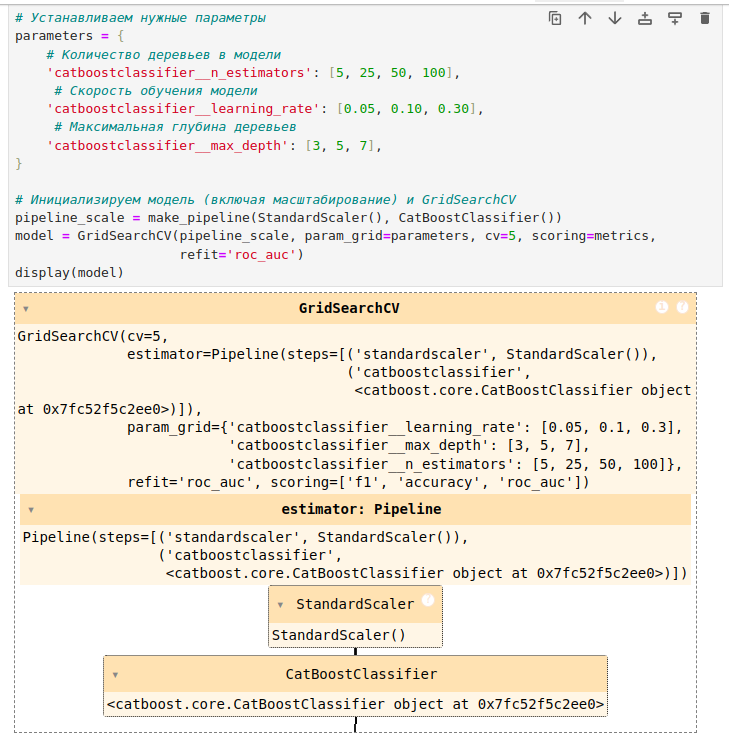
# LogisticRegression



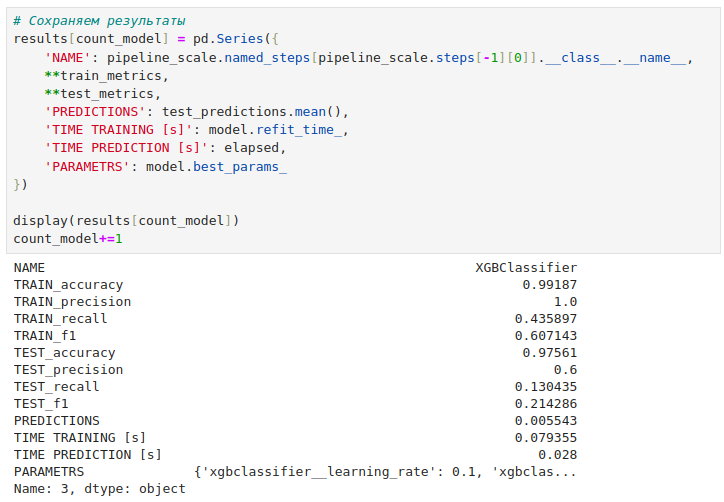
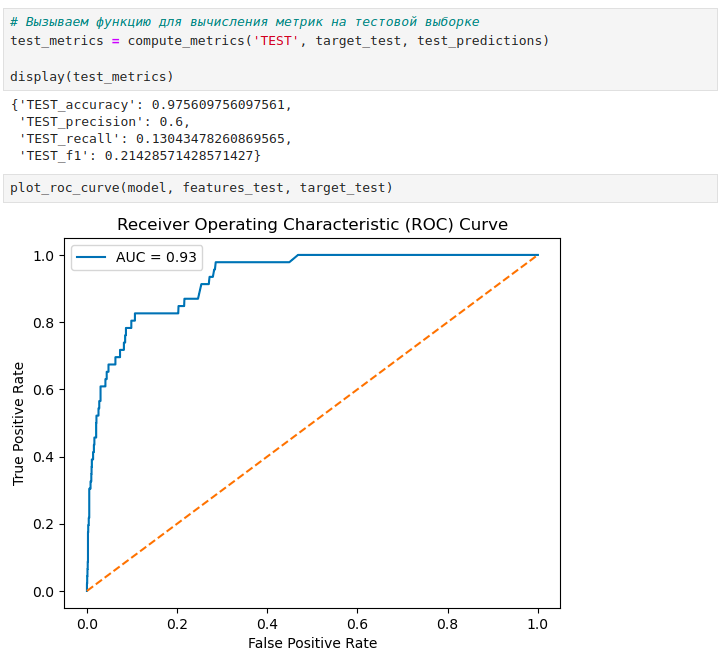
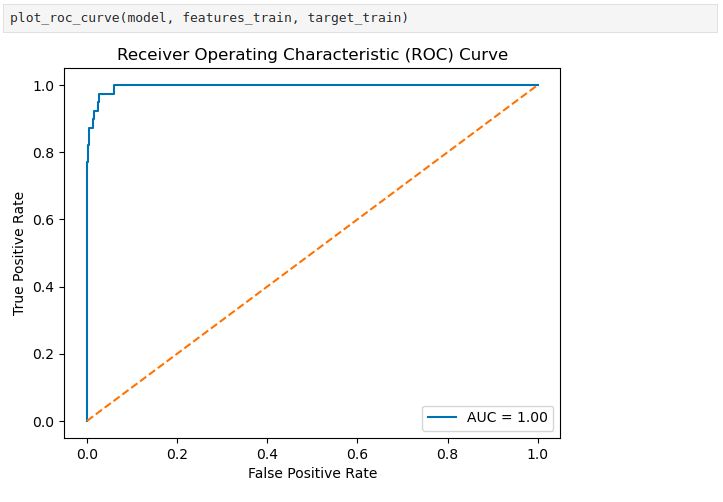
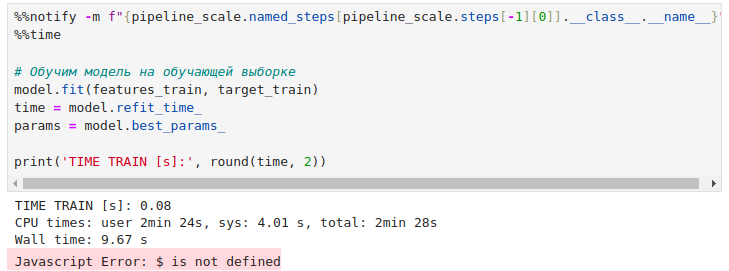
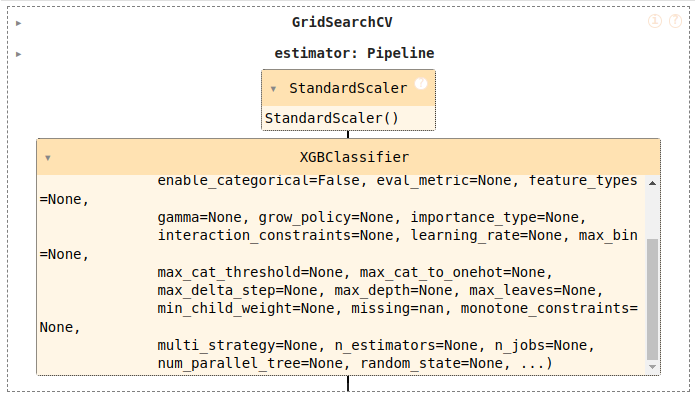
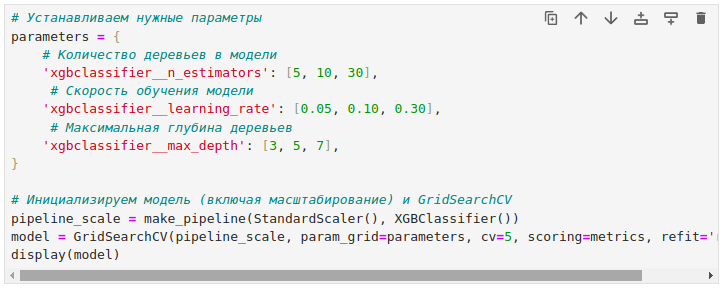
# LGBMClassifier



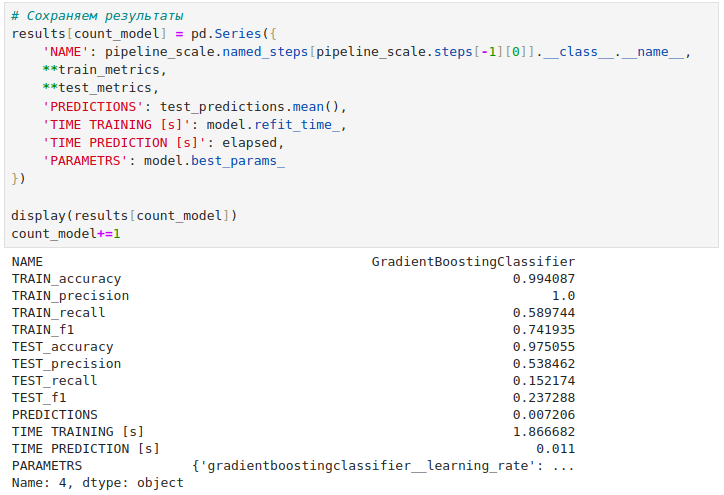
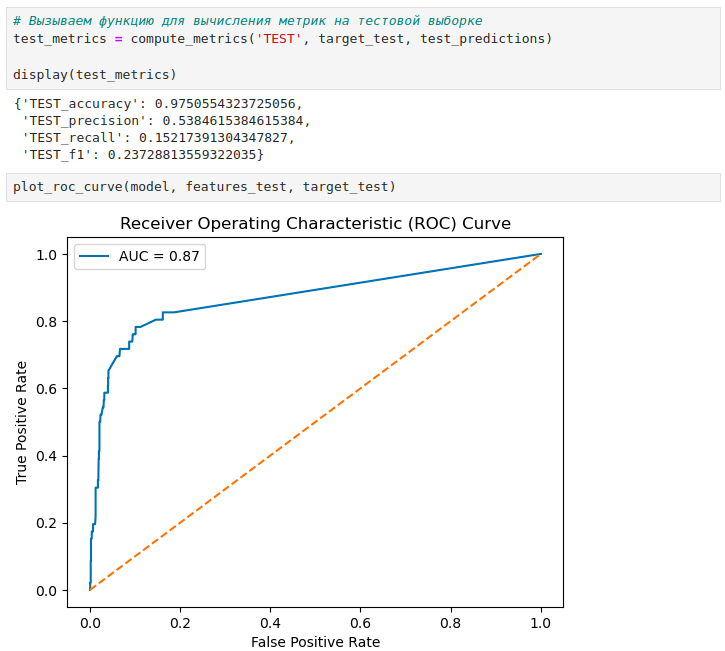
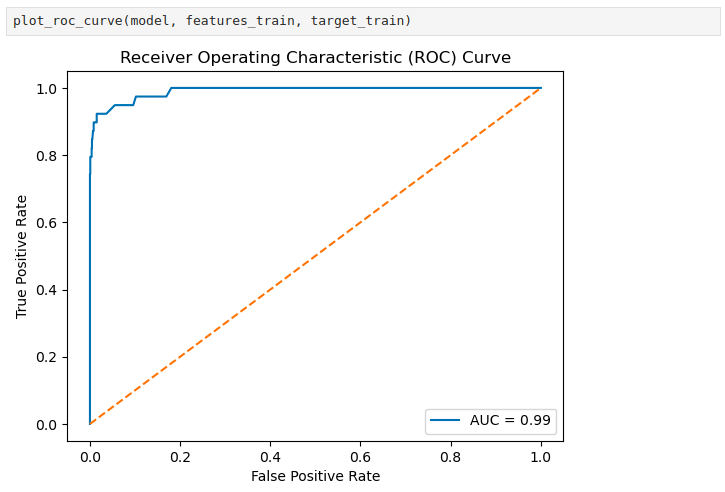
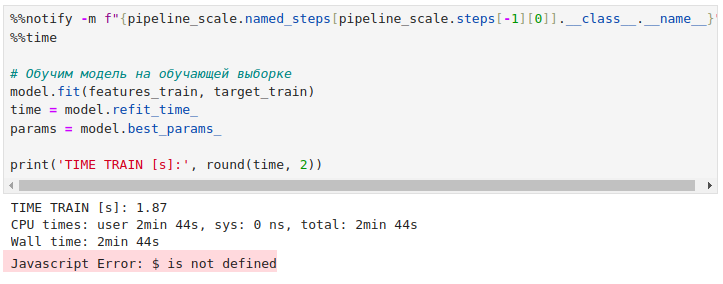
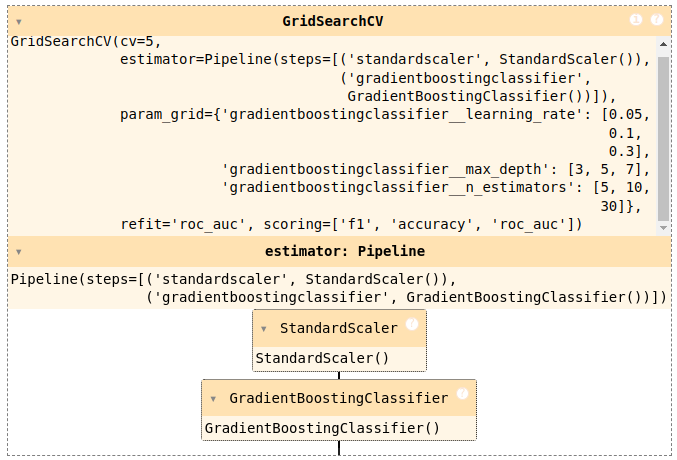
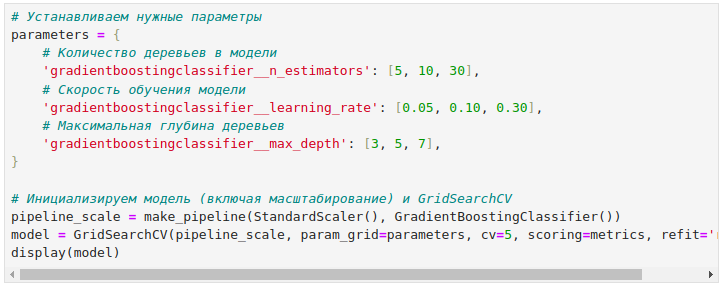
# CatBoostClassifier



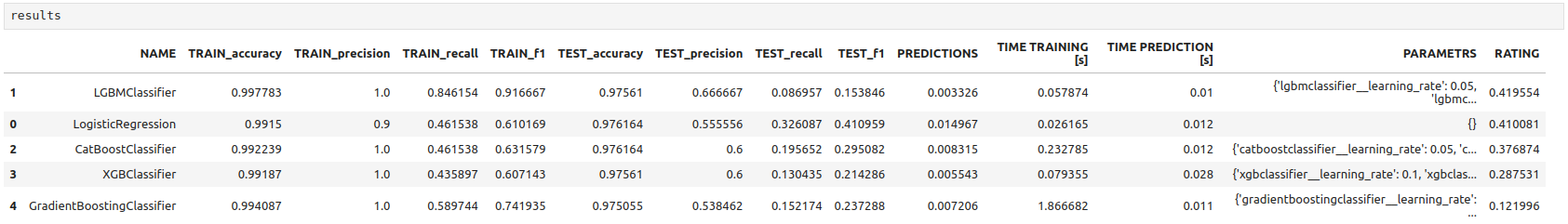
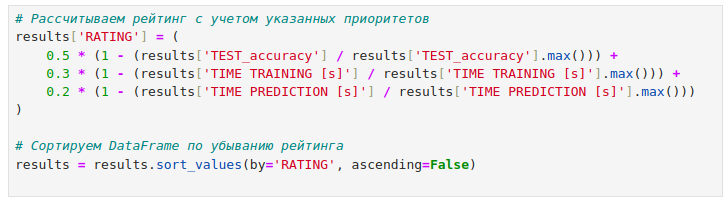
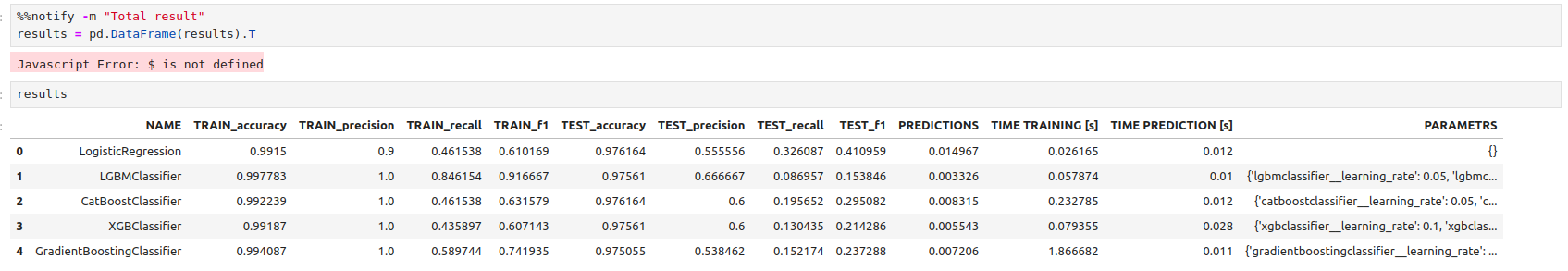
# XGBClassifier



# GradientBoostingClassifier



# Анализ моделей



* GradientBoostingClassifier: Показывает идеальную точность на обучающей выборке (1.0), но значительно худшие результаты на тестовой выборке (0.9684). Это может указывать на переобучение модели.
* LGBMClassifier: Имеет высокую точность на обеих выборках (0.9978 на обучающей и 0.9756 на тестовой). Показывает лучший результат по метрике recall на тестовой выборке (0.8462). Время обучения и прогнозирования относительно невелико.
* CatBoostClassifier: Также показывает высокую точность на обеих выборках (0.9922 на обучающей и 0.9762 на тестовой). Время обучения немного больше, чем у LGBMClassifier.
* LogisticRegression: Имеет неплохую точность на обеих выборках (0.9915 на обучающей и 0.9762 на тестовой). Показывает худшие результаты по метрикам precision и recall на тестовой выборке.
* XGBClassifier: Показывает самую низкую точность на тестовой выборке (0.9756) среди всех моделей, кроме GradientBoostingClassifier.

# Вывод

Удивительно, но LGBMClassifier показала себя лучшей по сравнению с остальными моделями. Да и не говоря о скорости обучения и предсказани. Второе место претендует LogisticRegression, третье место - CatBoostClassifier.

# Гистограмма

