**Московский государственный технический   
университет им. Н. Э. Баумана**

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №2

«Методы построения моделей машинного обучения»

Вариант № 31

Выполнил: Проверил:  
Падалко К.Р. Гапанюк Ю.Е.  
группа ИУ5Ц-84Б

Дата: 28.04.25 Дата:

Подпись: Подпись:

Москва, 2025 г.

Содержание

[1. Задание 3](#_Toc196750606)

[2. Листинг 4](#_Toc196750607)

[2.1. Подключение библиотеки и импорт данных 4](#_Toc196750608)

[2.2. Изучение данных 5](#_Toc196750609)

[2.3. Пропуски 6](#_Toc196750610)

[2.4. Дубликаты 6](#_Toc196750611)

[2.5. Обработка категориальных признаков 6](#_Toc196750612)

[2.6. Обучение и тестирование 6](#_Toc196750613)

[2.6.1. Метод 1: Логическая регрессия 7](#_Toc196750614)

[2.6.2. Метод 2: Градиентный бустинг 8](#_Toc196750615)

[2.6.3. Оценка качеств моделей 9](#_Toc196750616)

[2.7. Вывод 9](#_Toc196750617)

## 1. 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:**

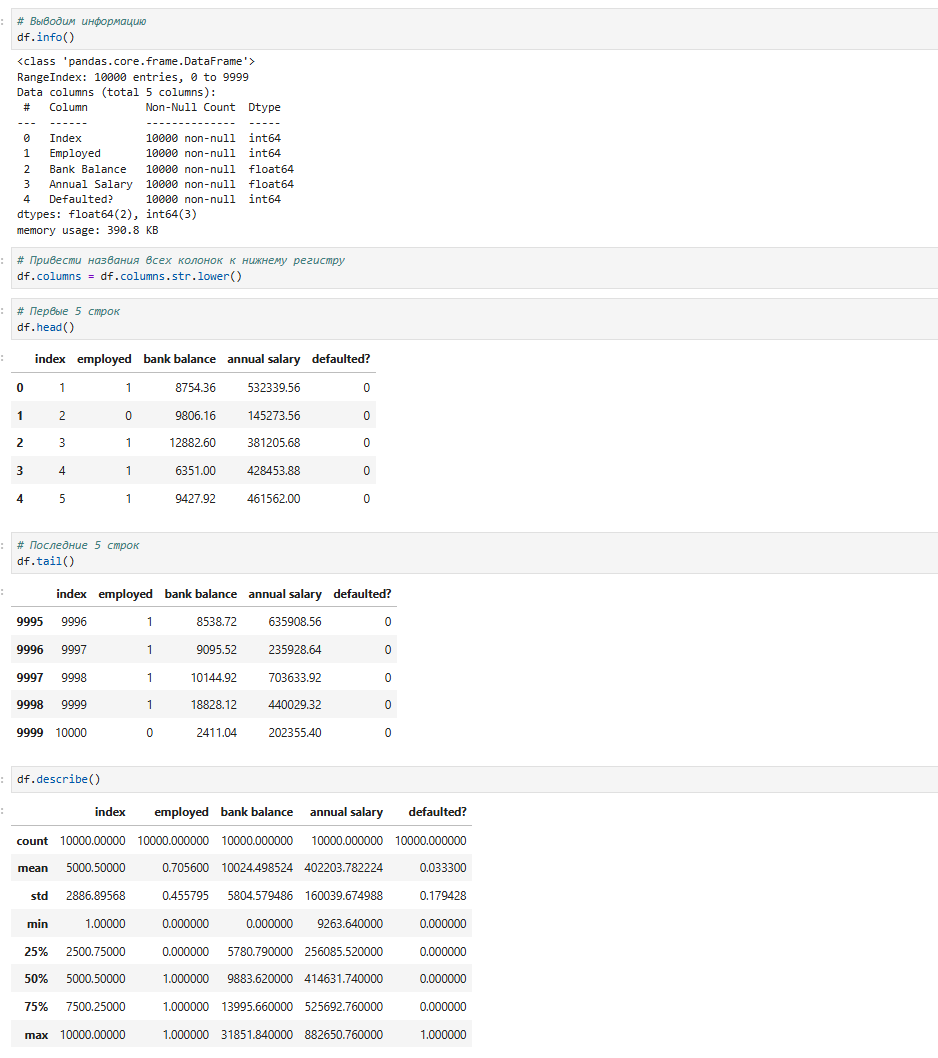
<https://www.kaggle.com/kmldas/loan-default-prediction>

## 2. Листинг

## 2.1. Подключение библиотеки и импорт данных



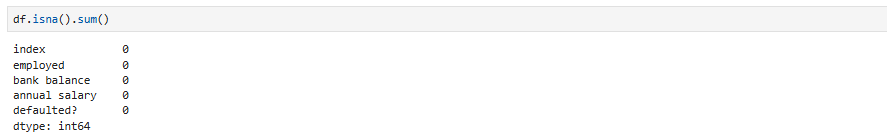
## 2.2. Изучение данных



Анализ описательной статистики:

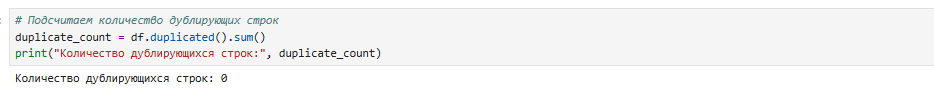
* employed: примерно 70% людей трудоустроены.
* bank balance: на банковском счете: большая часть людей имеет баланс на счете меньше 14 000.
* anual salary: существуют значительные различия в уровне зарплат, от минимальных до очень высоких (до почти 883 000).
* defaulted: только около 3% людей имеют дефолт по кредиту, что свидетельствует о сравнительно низком уровне риска дефолта в этой выборке.

## 2.3. Пропуски



В наборе нет пропусков, следовательно, не нужно их обрабатывать.

## 2.4. Дубликаты



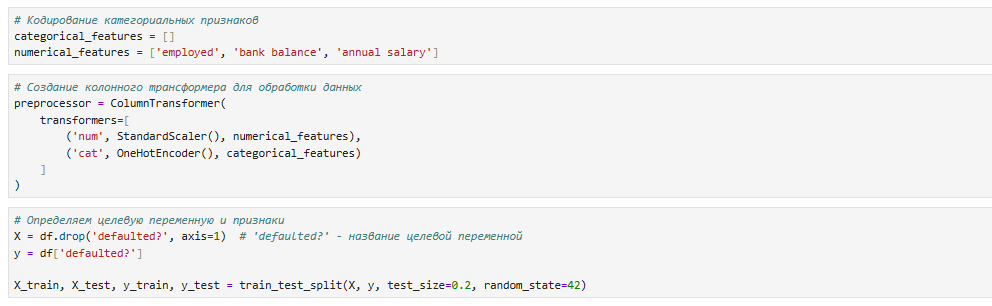
В наборе нет дублирующих значений, следовательно, не нужно их обрабатывать.

## 2.5. Обработка категориальных признаков



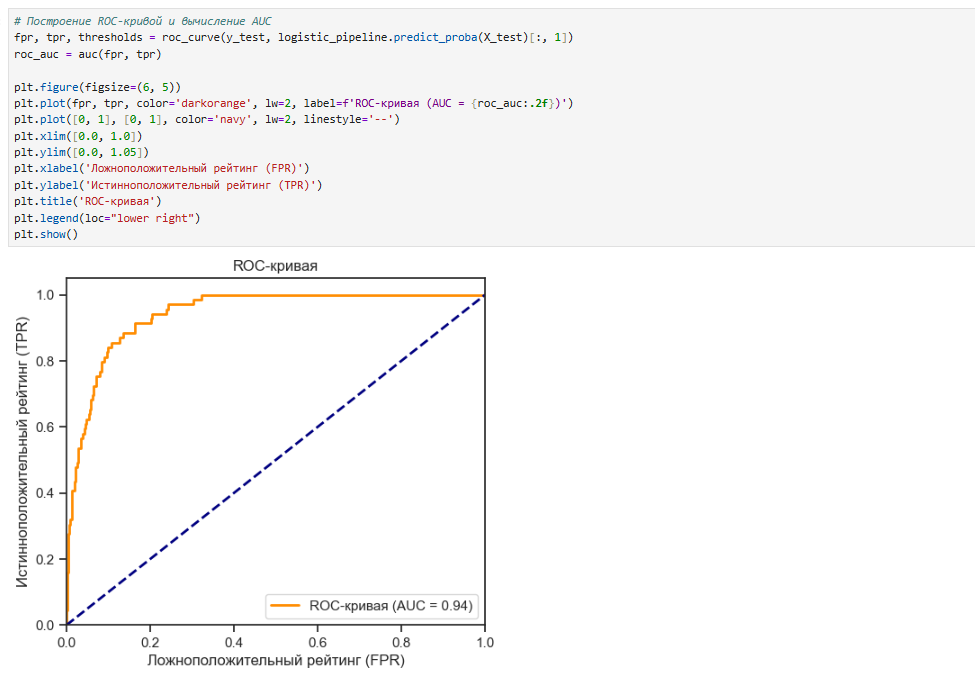
Нет категориальных значений, значит ненужно кодировать категориальных признаков.

## 2.6. Обучение и тестирование

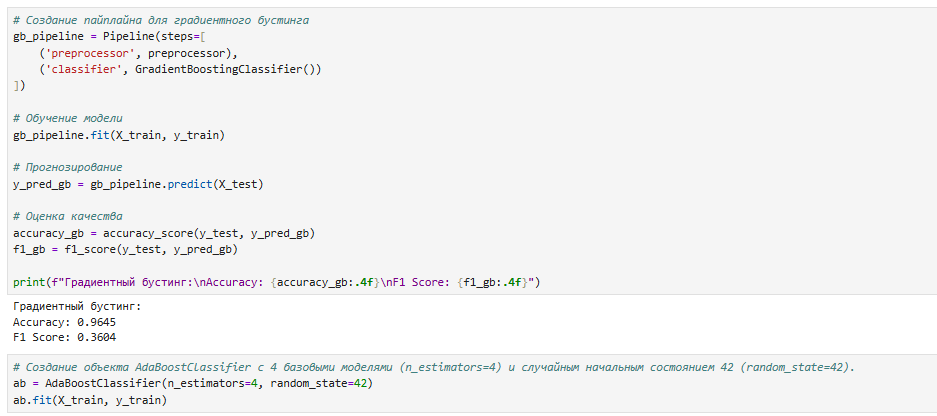


## 2.6.1. Метод 1: Логическая регрессия

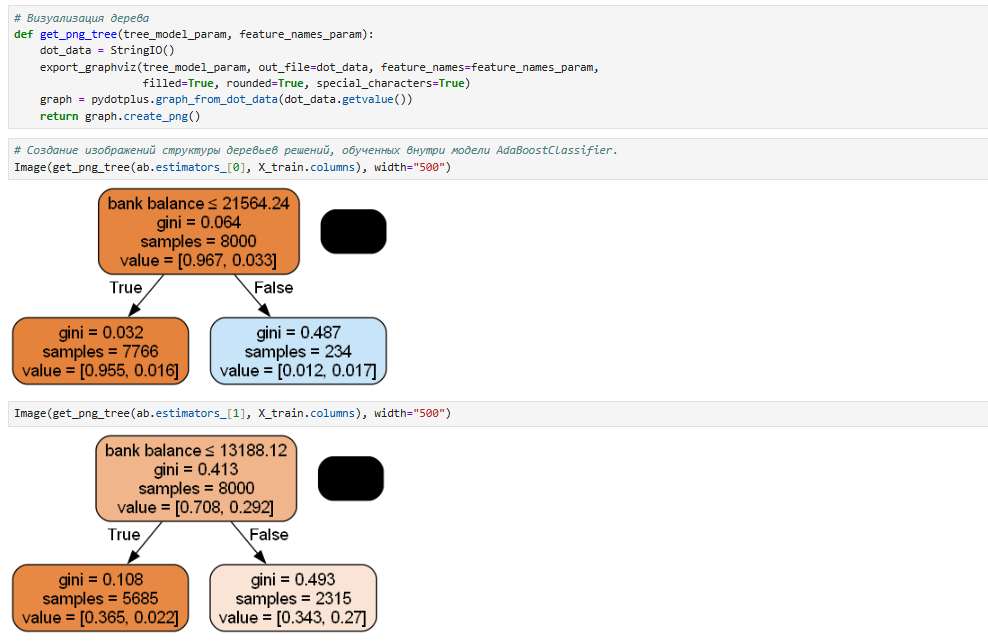




## 2.6.2. Метод 2: Градиентный бустинг

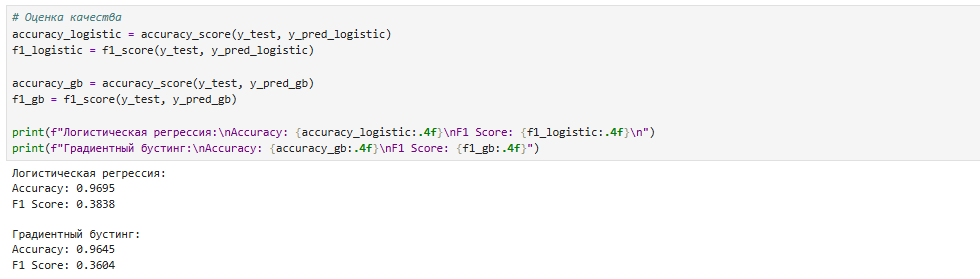








## 2.6.3. Оценка качеств моделей



## 2.7. Вывод

Градиентный бустинг имеет несколько более низкие значения Accuracy и F1 Score, чем Логистическая регрессия, что указывает на то, что модель градиентного бустинга лучше справляется с задачей классификации.

Логистическая регрессия показывает более высокую точность, но её F1 Score заметно ниже, что может свидетельствовать о проблемах с балансом классов или недостаточной способности модели захватывать сложные зависимости между признаками и целевой переменной.

Градиентный бустинг лучше справляется с захватом более сложных нелинейных взаимосвязей между признаками и целевой переменной, что позволяет достигать лучшего качества классификации в условиях, когда логистическая регрессия может быть ограничена своими предположениями о линейности модели.

Таким образом, при выборе модели для данной задачи предпочтительнее использовать градиентный бустинг, так как он предоставляет более сбалансированные результаты и может лучше учитывать сложные зависимости в данных.