МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ АСТАНА

Высшая школа информационных технологий и инженерии

ОТЧЕТ

по сессионному проекту

Дисциплина: Основы интеллектуального анализа данных и машинного обучения

Выполнили: студенты 3 курса специальности Data Science Испанов Ермек, Кульмагамбетов Бий

1. Введение и цель проекта

В рамках сессионного проекта по дисциплине «Машинное обучение» была выполнена практическая работа по построению и анализу моделей машинного обучения на реальном банковском датасете. Целью проекта являлось не только получение качественных предсказаний, но и глубокое понимание принципов работы базовых алгоритмов машинного обучения за счёт их реализации с нуля, без использования готовых моделей из библиотек.

В проекте решались две основные задачи:

- **Задача регрессии** — предсказание числового показателя баланса клиента (balance).
- **Задача классификации** — определение того, оформит ли клиент банковский депозит (deposit).

Для решения задач были реализованы и проанализированы следующие модели:

- линейная регрессия (реализация с нуля);
- логистическая регрессия (реализация с нуля);
- модель Random Forest (с использованием библиотеки sklearn для сравнения качества).

Отдельное внимание уделялось экспериментам с гиперпараметрами, анализу метрик качества и визуализации результатов.

2. Описание и источник данных

В проекте использовался реальный открытый датасет **Bank Marketing Dataset**, предоставленный на платформе Kaggle.

Источник данных: https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset

Датасет содержит информацию о клиентах банка и результатах маркетинговых кампаний по привлечению вкладчиков.

Основные характеристики датасета:

Количество наблюдений: 45 211 строк

Количество признаков: 17

Тип данных: структурированные табличные данные

Описание признаков (основные):

- age — возраст клиента
- job — тип занятости
- marital — семейное положение
- education — уровень образования
- balance — баланс счёта клиента
- housing — наличие ипотечного кредита
- loan — наличие потребительского кредита
- duration — длительность последнего контакта
- campaign — количество контактов в рамках кампании
- deposit — целевая переменная (оформил депозит или нет)

Для задачи регрессии использовалась переменная balance, а для задачи классификации — deposit.

3. Предобработка данных

Перед построением моделей была выполнена стандартная предобработка данных:

1. **Обработка категориальных признаков**
Категориальные переменные (job, marital, education и др.) были преобразованы в числовой формат с использованием кодирования.
2. **Выделение целевых переменных**
Для регрессии: y = balance
Для классификации: y = deposit
3. **Масштабирование признаков**
Числовые признаки были приведены к сопоставимым масштабам, что существенно ускоряет и стабилизирует работу градиентного спуска.

4. **Добавление свободного члена (intercept)**
Для линейной и логистической регрессии вручную был добавлен столбец единиц.

4. Линейная регрессия (реализация с нуля)

Линейная регрессия использовалась для предсказания баланса клиента.

4.1 Математическая модель

Модель линейной регрессии имеет вид:

$y = wX + b$

В качестве функции потерь использовалась среднеквадратичная ошибка (MSE):

$MSE = (1 / n) * \Sigma(y\_true - y\_pred)^2$

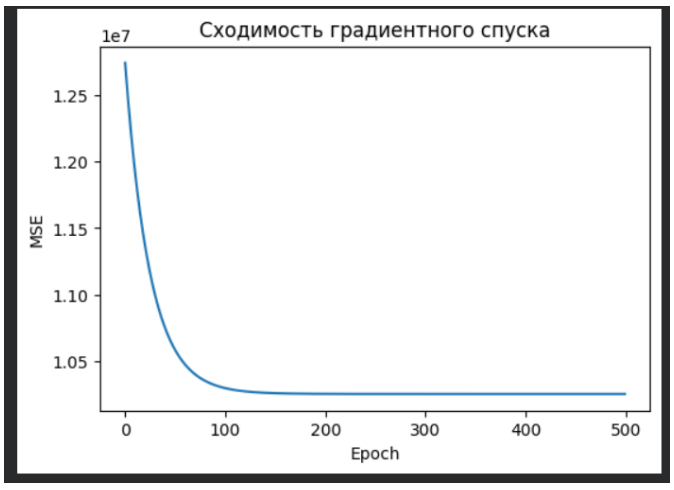
4.2 Обучение модели

Обучение модели осуществлялось с помощью batch gradient descent, реализованного вручную с использованием библиотеки NumPy. На каждой эпохе вычислялись градиенты по весам и свободному члену, после чего параметры обновлялись.

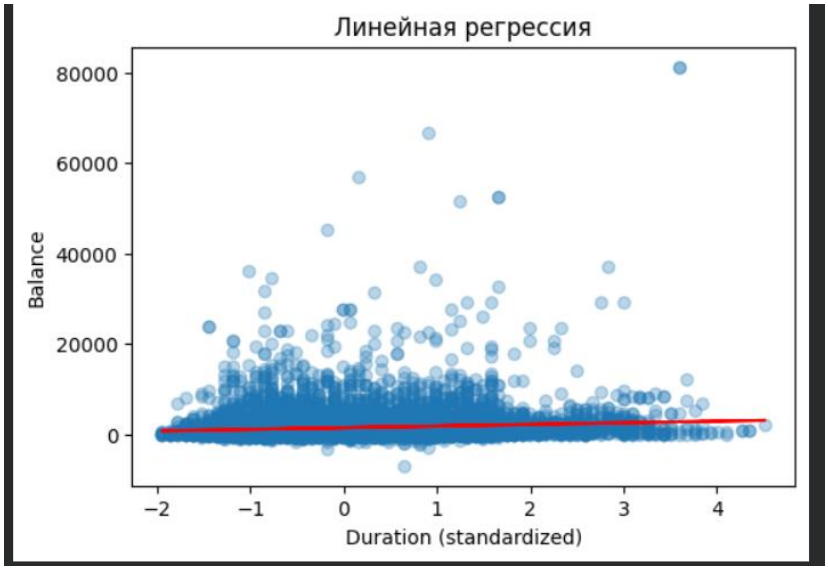
4.3 Результаты

В процессе обучения были построены:

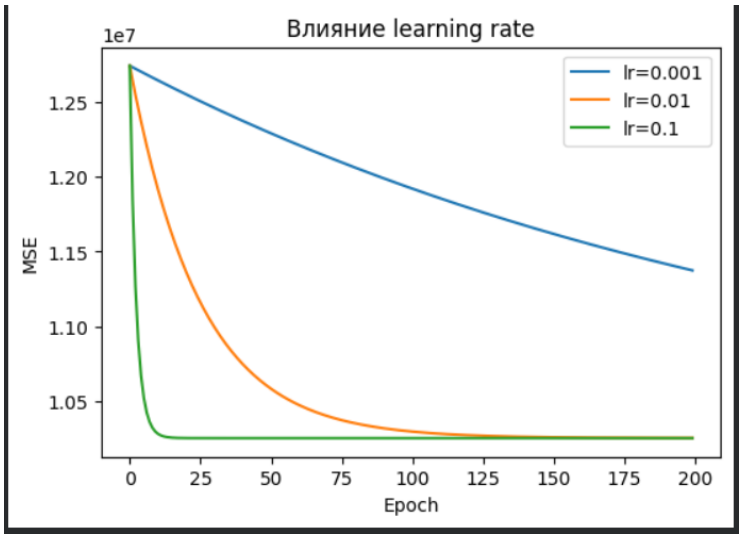
график функции потерь по эпохам;



scatter plot реальных данных и линия регрессии;



анализ влияния learning rate на скорость сходимости.



Модель успешно сошлась, что подтверждается монотонным уменьшением значения функции потерь.

5. Логистическая регрессия (реализация с нуля)

Логистическая регрессия применялась для бинарной классификации клиентов по признаку оформления депозита.

5.1 Математическая модель

В основе модели лежит сигмоидная функция:

$\sigma(z) = 1 / (1 + e^{-z})$

Функция потерь — log loss:

$L = -[y \cdot \log(p) + (1-y) \cdot \log(1-p)]$

5.2 Обучение модели

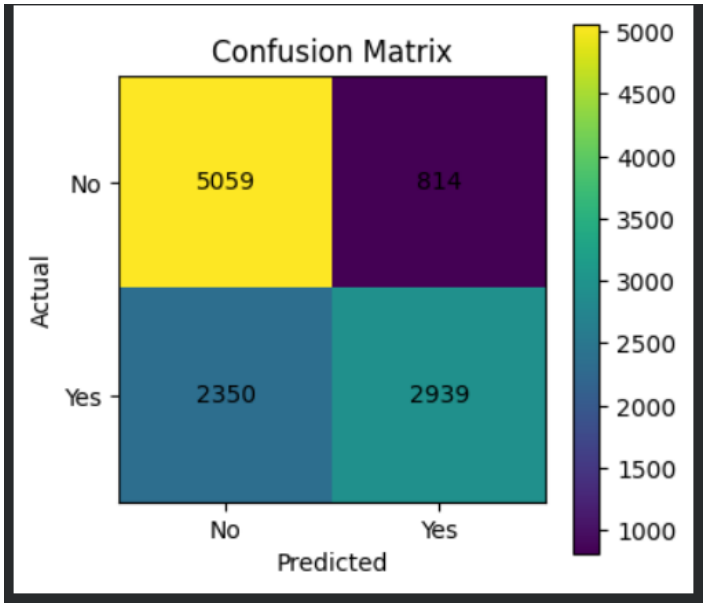
Как и в линейной регрессии, обучение выполнялось методом градиентного спуска, реализованного с нуля. На каждой итерации вычислялись градиенты функции потерь и обновлялись веса модели.

5.3 Метрики качества

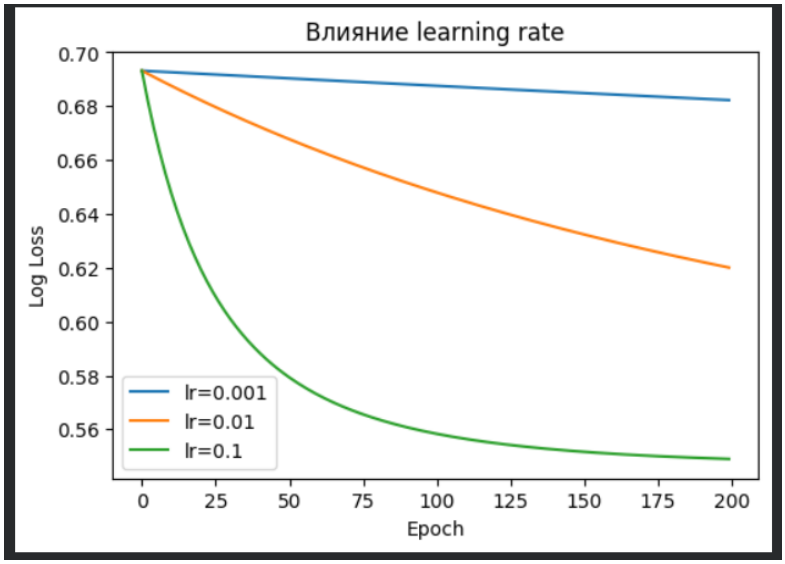
Для оценки классификации использовались:

- Accuracy
- Precision
- Recall
- F1-score
- ROC AUC
- Confusion Matrix

```
Accuracy: 0.7165382547930479
Precision: 0.7831068478550492
Recall: 0.5556816033276611
F1-score: 0.6500774165007742
ROC AUC: 0.8135217269991334
```



Были проведены эксперименты с разными значениями learning rate и количеством эпох, что позволило оценить устойчивость модели.



6. Random Forest и сравнение моделей

В качестве альтернативной модели классификации был использован Random Forest из библиотеки sklearn.

6.1 Причина выбора модели

Random Forest позволяет:

учитывать нелинейные зависимости; быть менее чувствительным к масштабу признаков; снижать переобучение за счёт ансамблирования.

6.2 Сравнение результатов

Результаты логистической регрессии и Random Forest были сравнены по всем основным метрикам. В большинстве случаев Random Forest показал более высокое качество классификации, что объясняется его способностью моделировать сложные зависимости между признаками.

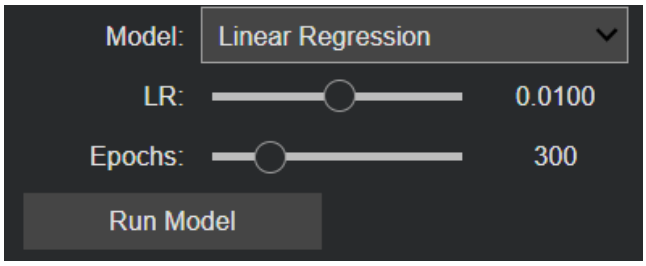
	Model	Accuracy	F1-score	ROC AUC
0	Logistic Regression	0.714734	0.653616	0.812505
1	Random Forest	0.757277	0.753412	0.841476

7. Эксперименты и визуализация

В рамках проекта были проведены эксперименты:

- влияние learning rate на скорость и стабильность сходимости;
- изменение количества эпох обучения;
- анализ поведения функции потерь.

Для удобства демонстрации был реализован интерактивный интерфейс в Google Colab, позволяющий выбирать модель и изменять параметры обучения в реальном времени.



8. Выводы

В ходе выполнения проекта были реализованы и исследованы методы машинного обучения на основе реального банковского датасета. Линейная и логистическая регрессии были реализованы с нуля, что позволило подробно изучить работу градиентного спуска, влияние скорости обучения и количества эпох на сходимость моделей. Эксперименты показали, что корректный выбор гиперпараметров напрямую влияет на качество обучения и стабильность алгоритмов.

Для задачи классификации была реализована логистическая регрессия и проведено сравнение с моделью Random Forest. Результаты показали, что логистическая регрессия является простой и интерпретируемой базовой моделью, однако Random Forest демонстрирует более высокое качество классификации за счёт способности учитывать сложные и нелинейные зависимости между признаками. В результате цели проекта были достигнуты, а полученные результаты подтвердили эффективность применения различных моделей машинного обучения для анализа банковских данных.