

Выполнил: Тимошинов Егор Борисович

Группа: 16

Лабораторная работа

Деревья решений для классификации

Цель работы

Изучение методов построения деревьев решений для решения задач классификации. Построение дерева решений для предсказания возврата кредита на основе характеристик клиента.

Задание

- 1. Подготовить данные о клиентах банка для построения дерева решений. Выполнить исследовательский анализ данных и визуализацию распределения признаков.
- 2. Построить дерево решений для классификации возврата кредита с использованием алгоритма CART.
- 3. Проанализировать важность признаков в построенном дереве решений.
- 4. Выполнить классификацию на тестовой выборке и построить таблицу сопряженности.
- 5. Оценить точность модели с помощью различных метрик классификации.

Результаты выполнения задания

Задание 1. Подготовка и анализ данных

Для построения модели использовался набор данных о клиентах банка, включающий следующие признаки:

- Возраст клиента (количественный признак)
- Доход клиента (количественный признак)
- Кредитная история (категориальный признак: хорошая, средняя, плохая)
- Цель кредита (категориальный признак: потребительский, автомобиль, недвижимость, бизнес)
- Срок кредита в месяцах (количественный признак)

Целевая переменная: возврат кредита (Да/Нет).

Общее количество наблюдений: 1000. Данные разделены на обучающую выборку (70%) и тестовую выборку (30%).

Описательная статистика количественных признаков:

Признак	Среднее	Стандартное отклонение	Минимум	Максимум
---------	---------	------------------------	---------	----------

Возраст	43.82	14.99	18	69
Доход	84905.98	38430.89	20060	149972
Срок кредита	36.14	17.04	12	60

На рисунках ниже представлены распределения признаков и целевой переменной:

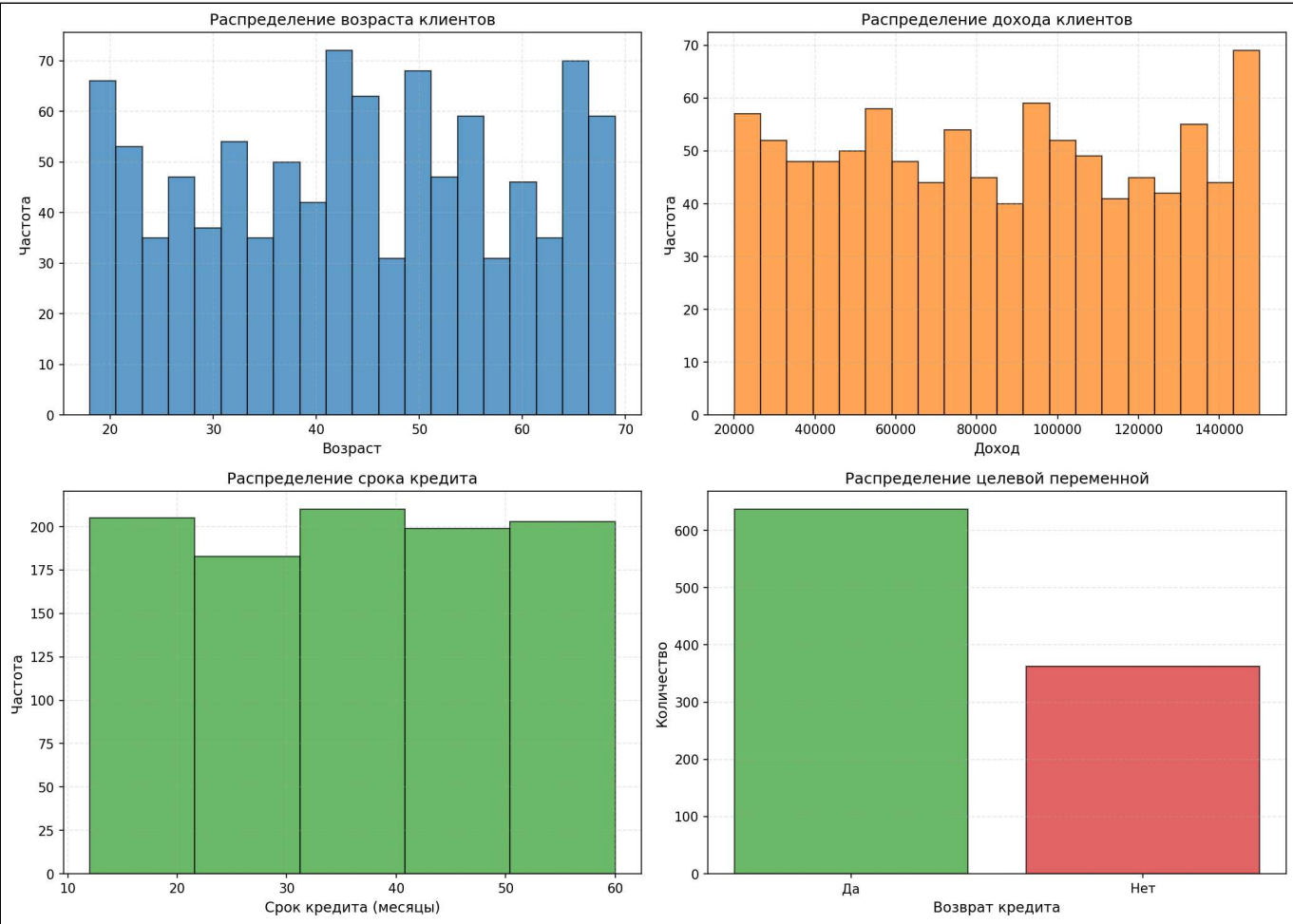


Рисунок 1. Распределение количественных признаков и целевой переменной

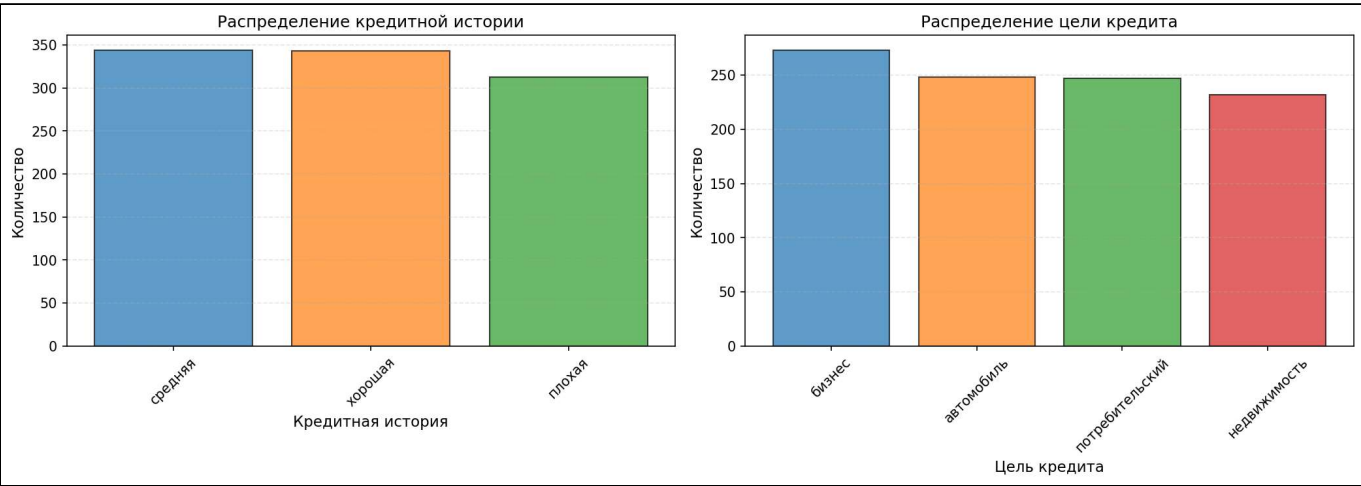


Рисунок 2. Распределение категориальных признаков

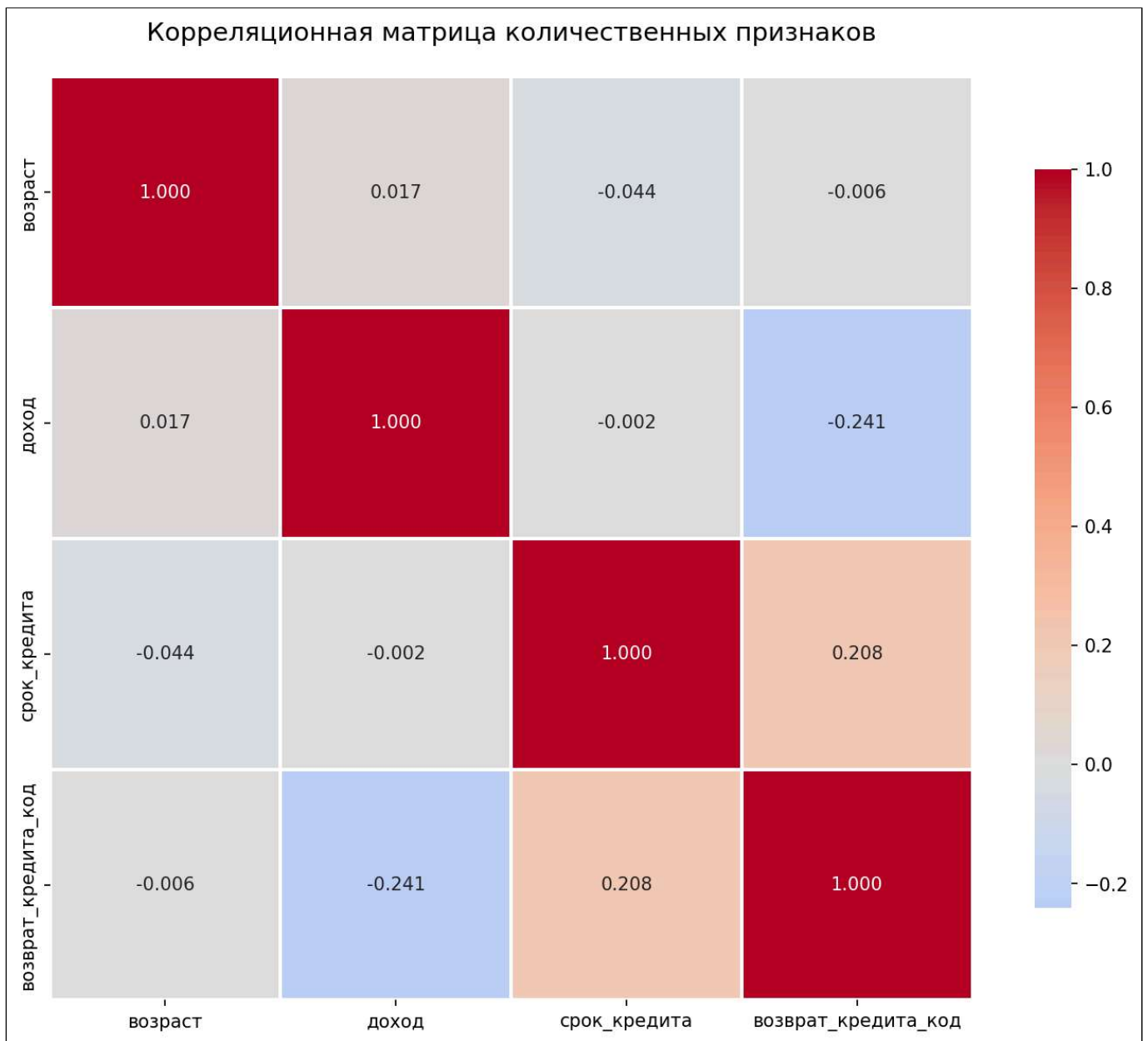


Рисунок 3. Корреляционная матрица количественных признаков

Задание 2. Построение дерева решений

Для построения дерева решений использовался алгоритм CART (Classification and Regression Trees) с критерием разделения Gini. Параметры модели:

- Максимальная глубина дерева: 5
- Минимальное количество образцов для разделения узла: 20
- Минимальное количество образцов в листе: 5

Построенное дерево решений представлено на рисунке ниже.

Дерево решений для классификации возврата кредита

```

graph TD
    Root["Кредитная история <= 0.5  
gini = 0.47  
samples = 700  
value = [436, 264]  
class = Нет"]
    
    Root -- True --> Node1["Доход <= 84603.5  
gini = 0.471  
samples = 221  
value = [84, 137]  
class = Да"]
    Root -- False --> Node2["Доход <= 48817.0  
gini = 0.39  
samples = 479  
value = [352, 127]  
class = Нет"]
    
    Node1 --> Node3["Срок кредита <= 30.0  
gini = 0.365  
samples = 104  
value = [25, 79]  
class = Да"]
    Node1 --> Node4["Срок кредита <= 30.0  
gini = 0.5  
samples = 117  
value = [59, 58]  
class = Нет"]
    
    Node2 --> Node5["Кредитная история <= 1.5  
gini = 0.5  
samples = 117  
value = [59, 58]  
class = Нет"]
    Node2 --> Node6["Срок кредита <= 42.0  
gini = 0.309  
samples = 362  
value = [293, 69]  
class = Нет"]
    
    Node3 --> Node7["Доход <= 39551.5  
gini = 0.484  
samples = 34  
value = [14, 20]  
class = Да"]
    Node3 --> Node8["Цель кредита <= 1.5  
gini = 0.265  
samples = 70  
value = [11, 59]  
class = Да"]
    
    Node4 --> Node9["Возраст <= 34.5  
gini = 0.439  
samples = 40  
value = [27, 13]  
class = Нет"]
    Node4 --> Node10["Возраст <= 55.5  
gini = 0.486  
samples = 77  
value = [32, 45]  
class = Да"]
    
    Node5 --> Node11["Возраст <= 66.5  
gini = 0.476  
samples = 69  
value = [27, 42]  
class = Да"]
    Node5 --> Node12["Возраст <= 48.0  
gini = 0.444  
samples = 48  
value = [32, 16]  
class = Нет"]
    
    Node6 --> Node13["Доход <= 75393.5  
gini = 0.179  
samples = 211  
value = [190, 21]  
class = Нет"]
    Node6 --> Node14["Кредитная история <= 1.5  
gini = 0.434  
samples = 151  
value = [103, 48]  
class = Нет"]
    
    Node7 --> Node15["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    Node7 --> Node16["Доход <= 67180.5  
gini = 0.097  
samples = 39  
value = [2, 37]  
class = Да"]
    
    Node8 --> Node17["Доход <= 41000.0  
gini = 0.4  
samples = 9  
value = [9]  
class = Да"]
    Node8 --> Node18["Срок кредита <= 0.5  
gini = 0  
samples = 61  
value = [0]  
class = Да"]
    
    Node9 --> Node19["Возраст <= 52.5  
gini = 0.49  
samples = 28  
value = [16, 12]  
class = Нет"]
    Node9 --> Node20["Доход <= 100786.5  
gini = 0.5  
samples = 55  
value = [27.0, 28.0]  
class = Да"]
    
    Node10 --> Node21["Возраст <= 62.5  
gini = 0.351  
samples = 22  
value = [5, 17]  
class = Да"]
    Node10 --> Node22["Возраст <= 55.5  
gini = 0.476  
samples = 69  
value = [32, 16]  
class = Нет"]
    
    Node11 --> Node23["Возраст <= 66.5  
gini = 0.476  
samples = 69  
value = [27, 42]  
class = Да"]
    Node11 --> Node24["Цель кредита <= 0.5  
gini = 0  
samples = 2  
value = [2]  
class = Да"]
    
    Node12 --> Node25["Кредитная история <= 1.5  
gini = 0.367  
samples = 62  
value = [47, 15]  
class = Нет"]
    Node12 --> Node26["Возраст <= 51.5  
gini = 0.077  
samples = 149  
value = [143, 6]  
class = Нет"]
    
    Node13 --> Node27["Возраст <= 55.5  
gini = 0.496  
samples = 77  
value = [42, 35]  
class = Нет"]
    Node13 --> Node28["Возраст <= 52.0  
gini = 0.29  
samples = 74  
value = [61, 13]  
class = Нет"]
    
    Node15 --> Node29["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    Node15 --> Node30["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    
    Node16 --> Node31["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    Node16 --> Node32["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    
    Node17 --> Node33["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    Node17 --> Node34["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    
    Node18 --> Node35["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    Node18 --> Node36["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    
    Node19 --> Node37["Возраст <= 52.5  
gini = 0.49  
samples = 28  
value = [16, 12]  
class = Нет"]
    Node19 --> Node38["Возраст <= 52.5  
gini = 0.49  
samples = 28  
value = [16, 12]  
class = Нет"]
    
    Node20 --> Node39["Возраст <= 52.5  
gini = 0.49  
samples = 28  
value = [16, 12]  
class = Нет"]
    Node20 --> Node40["Возраст <= 52.5  
gini = 0.49  
samples = 28  
value = [16, 12]  
class = Нет"]
    
    Node21 --> Node41["Возраст <= 62.5  
gini = 0.351  
samples = 22  
value = [5, 17]  
class = Да"]
    Node21 --> Node42["Возраст <= 62.5  
gini = 0.351  
samples = 22  
value = [5, 17]  
class = Да"]
    
    Node22 --> Node43["Возраст <= 55.5  
gini = 0.476  
samples = 69  
value = [32, 16]  
class = Нет"]
    Node22 --> Node44["Возраст <= 55.5  
gini = 0.476  
samples = 69  
value = [32, 16]  
class = Нет"]
    
    Node23 --> Node45["Возраст <= 66.5  
gini = 0.476  
samples = 69  
value = [27, 42]  
class = Да"]
    Node23 --> Node46["Возраст <= 66.5  
gini = 0.476  
samples = 69  
value = [27, 42]  
class = Да"]
    
    Node24 --> Node47["Цель кредита <= 0.5  
gini = 0  
samples = 2  
value = [2]  
class = Да"]
    Node24 --> Node48["Цель кредита <= 0.5  
gini = 0  
samples = 2  
value = [2]  
class = Да"]
    
    Node25 --> Node49["Кредитная история <= 1.5  
gini = 0.367  
samples = 62  
value = [47, 15]  
class = Нет"]
    Node25 --> Node50["Кредитная история <= 1.5  
gini = 0.367  
samples = 62  
value = [47, 15]  
class = Нет"]
    
    Node26 --> Node51["Возраст <= 51.5  
gini = 0.077  
samples = 149  
value = [143, 6]  
class = Нет"]
    Node26 --> Node52["Возраст <= 51.5  
gini = 0.077  
samples = 149  
value = [143, 6]  
class = Нет"]
    
    Node27 --> Node53["Возраст <= 55.5  
gini = 0.496  
samples = 77  
value = [42, 35]  
class = Нет"]
    Node27 --> Node54["Возраст <= 55.5  
gini = 0.496  
samples = 77  
value = [42, 35]  
class = Нет"]
    
    Node28 --> Node55["Возраст <= 52.0  
gini = 0.29  
samples = 74  
value = [61, 13]  
class = Нет"]
    Node28 --> Node56["Возраст <= 52.0  
gini = 0.29  
samples = 74  
value = [61, 13]  
class = Нет"]
    
    Node29 --> Node57["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    Node29 --> Node58["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    
    Node30 --> Node59["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    Node30 --> Node60["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    
    Node31 --> Node61["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    Node31 --> Node62["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    
    Node32 --> Node63["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    Node32 --> Node64["Возраст <= 36.5  
gini = 0.49  
samples = 21  
value = [12, 9]  
class = Нет"]
    
    Node
```

Рисунок 4. Визуализация дерева решений

Задание 3. Анализ важности признаков

Важность признаков в построенном дереве решений:

Признак	Важность
Кредитная история	0.4131
Доход	0.2740
Возраст	0.1637
Срок кредита	0.1203
Цель кредита	0.0289

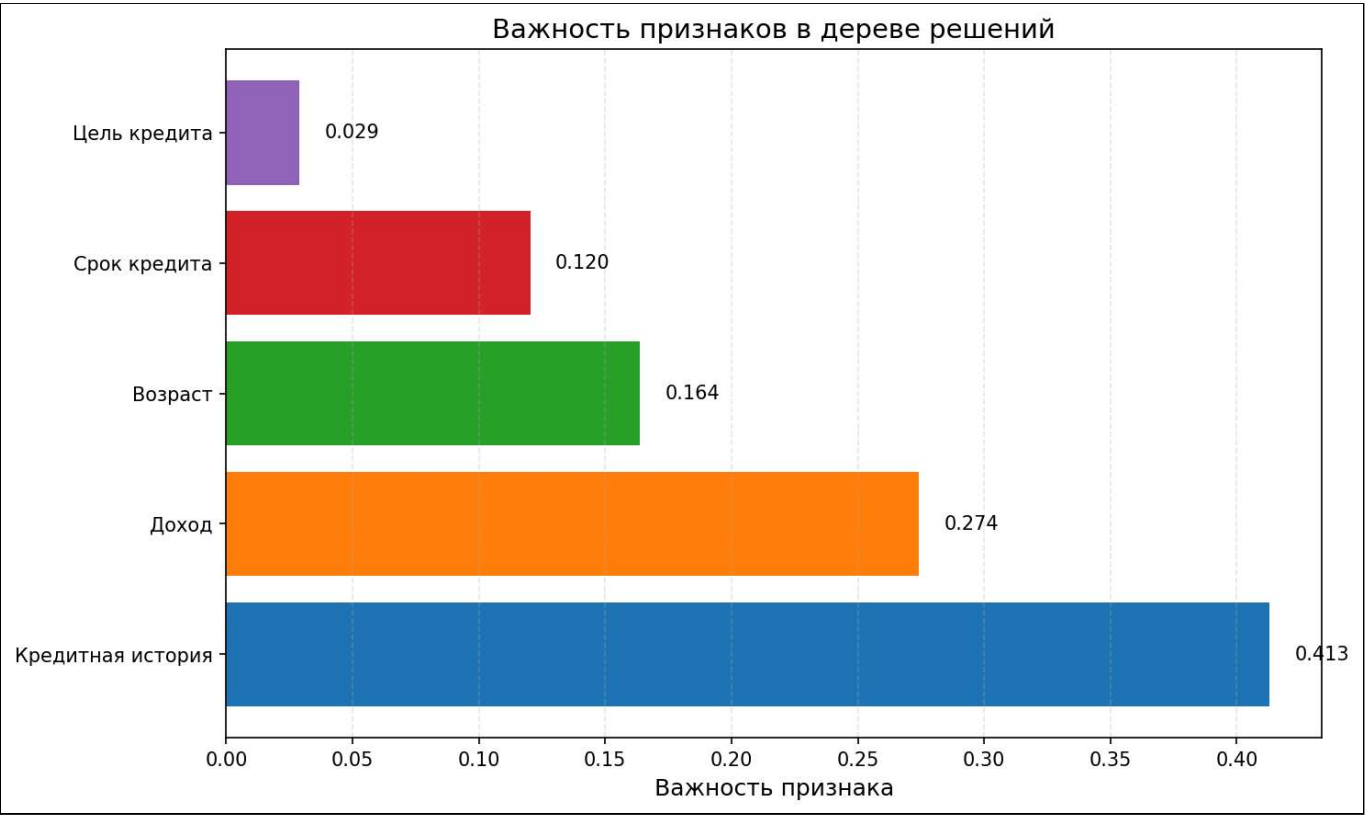


Рисунок 5. Важность признаков в дереве решений

Задание 4. Классификация на тестовой выборке

Для оценки качества построенной модели была выполнена классификация на тестовой выборке, содержащей 300 наблюдений.

Задание 5. Таблица сопряженности и оценка точности

Таблица сопряженности показывает количество правильно и неправильно классифицированных наблюдений:

	Предсказано: Нет	Предсказано: Да
Истинно: Нет	144	57
Истинно: Да	35	64

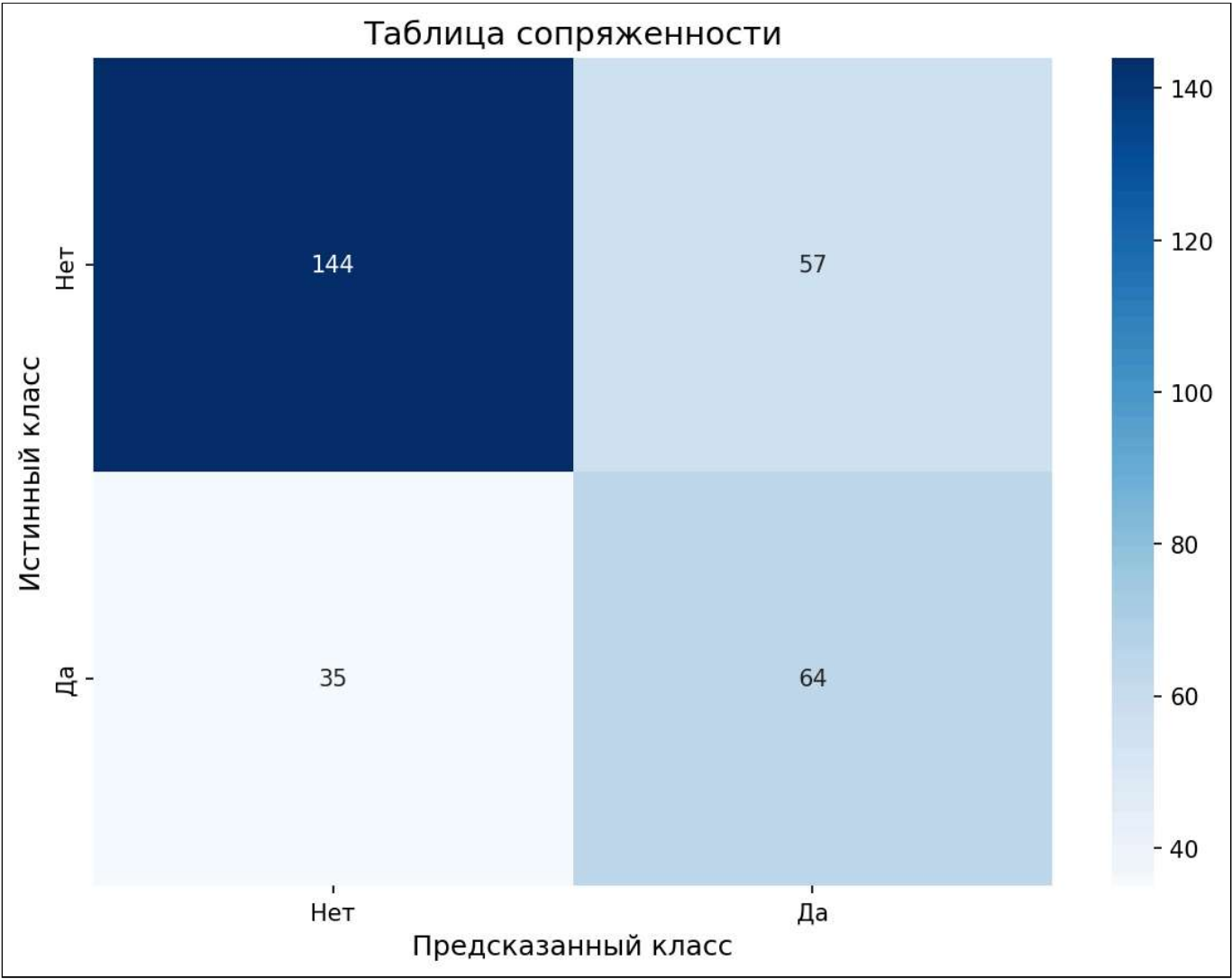


Рисунок 6. Таблица сопряженности

Характеристики точности построенной модели:

Метрика	Значение
Точность (Accuracy)	0.6933 (69.33%)
Правильно классифицировано	208 из 300
Ошибок классификации	92 из 300

Детальный отчет о классификации

Класс	Precision	Recall	F1-score	Support
Нет	0.8045	0.7164	0.7579	201
Да	0.5289	0.6465	0.5818	99
Среднее (macro avg)	0.6667	0.6814	0.6699	300

Среднее (weighted avg)	0.7135	0.6933	0.6998	300
------------------------	--------	--------	--------	-----

Текстовая структура дерева

Ниже представлена текстовая структура построенного дерева решений (первые уровни):

```
|--- Кредитная история <= 0.50
|   |--- Доход <= 84603.50
|   |   |--- Срок кредита <= 30.00
|   |   |   |--- Доход <= 39551.50
|   |   |   |   |--- class: 1
|   |   |   |   |--- Доход > 39551.50
|   |   |   |   |   |--- Возраст <= 36.50
|   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- Возраст > 36.50
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |--- Срок кредита > 30.00
|   |   |   |   |--- Цель кредита <= 1.50
|   |   |   |   |   |--- Доход <= 67180.50
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- Доход > 67180.50
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- Цель кредита > 1.50
|   |   |   |   |   |   |   |   |--- Доход <= 42164.50
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- Доход > 42164.50
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |--- Доход > 84603.50
|   |   |   |--- Срок кредита <= 30.00
|   |   |   |   |--- Возраст <= 34.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- Возраст > 34.50
|   |   |   |   |   |   |--- Возраст <= 52.50
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- Возраст > 52.50
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |--- Срок кредита > 30.00
|   |   |   |   |--- Возраст <= 55.50
|   |   |   |   |   |--- Доход <= 100786.50
|   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- Доход > 100786.50
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- Возраст > 55.50
|   |   |   |   |   |   |   |   |--- Возраст <= 62.50
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- Возраст > 62.50
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|--- Кредитная история > 0.50
|   |--- Доход <= 48817.00
|   |   |--- Кредитная история <= 1.50
|   |   |   |--- Возраст <= 66.50
|   |   |   |   |--- Возраст <= 53.00
|   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- Возраст > 53.00
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- Возраст > 66.50
|   |   |   |   |   |   |--- class: 0
|   |   |--- Кредитная история > 1.50
|   |   |   |--- Возраст <= 48.00
|   |   |   |   |--- Цель кредита <= 1.50
|   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- Цель кредита > 1.50
|   |   |   |   |   |   |--- class:...
```

Выводы

В ходе выполнения лабораторной работы было построено дерево решений для классификации возврата кредита. Модель показала точность 69.33% на тестовой выборке. Наиболее важными признаками для классификации оказались: Кредитная история (важность 0.4131) и Доход (важность 0.2740).

Из 300 наблюдений в тестовой выборке модель правильно классифицировала 208 наблюдений. Таблица сопряженности показывает, что модель лучше предсказывает класс "Нет" (не возврат кредита), чем класс "Да" (возврат кредита).

Построенное дерево решений может быть использовано банком для оценки риска выдачи кредита новым клиентам на основе их характеристик.