Отчет по лабораторной работе №1: Разведочный анализ данных (EDA) и подготовка к моделированию

Выполнили: студенты 6303-020302D Хасанов Дамир, Клейменов Андрей, Мергалиев Радмир, Павлов Виктор, Сидоров Артемий, Фокин Евгений.

Цель работы: Научиться загружать данные, выявлять проблемы (пропуски, выбросы, типы данных), строить понятные визуализации, формулировать гипотезы и готовить датасет к моделированию на примере набора данных "Маркетинг банковских услуг".

Глоссарий признаков датасета

Название колонки	Перевод и описание
age	Возраст клиента (числовой).
job	Профессия клиента (категориальный).
marital	Семейное положение (категориальный).
education	Уровень образования (категориальный).
default	Наличие кредитного дефолта (категориальный).
housing	Наличие ипотечного кредита (категориальный).
loan	Наличие потребительского кредита (категориальный).
contact	Тип связи при контакте (cellular/telephone).
month	Месяц последнего контакта.
day_of_week	День недели последнего контакта.
duration	Длительность последнего контакта в секундах.
campaign	Количество контактов с клиентом в рамках этой кампании.
pdays	Количество дней с момента прошлого контакта (999 - не контактировали).
previous	Количество контактов с клиентом до этой кампании.
poutcome	Результат предыдущей маркетинговой кампании.
emp.var.rate	Квартальный уровень изменения занятости в стране.
cons.price.idx	Месячный индекс потребительских цен (инфляция).
cons.conf.idx	Месячный индекс потребительского доверия.
euribor3m	Дневная ставка Euribor за 3 месяца.
nr.employed	Квартальное число занятых в стране.
y	Целевая переменная: согласился ли клиент на вклад (yes/no).

Этап 1: Первичный анализ и подготовка данных

1. Загрузка и осмотр: Был загружен датасет размером 41188 строк и 21 столбец. Первичный анализ показал наличие 10 числовых и 11 текстовых (object) признаков

```
--- 1.2. Первичный осмотр ---

Первые 5 строк:
```

	age	job	marital	education	default	housing	loan	contact	month	day_of_week
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon
1	57	services	married	high.school	unknown	no	no	telephone	may	mon
2	37	services	married	high.school	no	yes	no	telephone	may	mon
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon
4	56	services	married	high.school	no	no	yes	telephone	may	mon

5 rows × 21 columns

Последние 5 строк:

	age	job	marital	education	default	housing	loan	contact	month	day_
41183	73	retired	married	professional.course	no	yes	no	cellular	nov	
41184	46	blue-collar	married	professional.course	no	no	no	cellular	nov	
41185	56	retired	married	university.degree	no	yes	no	cellular	nov	
41186	44	technician	married	professional.course	no	no	no	cellular	nov	
41187	74	retired	married	professional.course	no	yes	no	cellular	nov	

5 rows × 21 columns

```
Размерность датасета (строки, столбцы): (41188, 21)
```

2. Проблемы качества данных:

```
Проверка качества данных

В датасете нет стандартных пропущенных значений (NaN).

Количество полностью дублирующихся строк: 12
Дубликаты удалены. Новая размерность датасета: (41176, 21)

Анализ и обработка 'unknown'
Значения 'unknown' являются скрытыми пропусками. Проанализируем их количество.
- Столбец 'job': 330 значений 'unknown' (0.80%)
- Столбец 'marital': 80 значений 'unknown' (0.19%)
- Столбец 'education': 1730 значений 'unknown' (4.20%)
- Столбец 'default': 8596 значений 'unknown' (20.88%)
- Столбец 'housing': 990 значений 'unknown' (2.40%)
- Столбец 'loan': 990 значений 'unknown' (2.40%)

Применяем стратегию: замена 'unknown' на самое частое значение (моду).
В столбце 'job' значения 'unknown' заменены на 'admin.'.
В столбце 'marital' значения 'unknown' заменены на 'married'.
В столбце 'education' значения 'unknown' заменены на 'university.degree'.
В столбце 'default' значения 'unknown' заменены на 'no'.
В столбце 'housing' значения 'unknown' заменены на 'yes'.
В столбце 'loan' значения 'unknown' заменены на 'no'.
```

   - Дубликаты: Обнаружено 12 полностью дублирующихся строк.

        - Решение: Дубликаты были удалены, так как они составляют менее 0.03% данных и, вероятнее всего, являются ошибками сбора информации.
    - "Скрытые" пропуски: Стандартных пропусков (NaN) не обнаружено. Однако в 6 категориальных столбцах были найдены значения unknown. Наибольшее их количество в столбце default (20.88%).
        - Решение: Вместо удаления строк, что привело бы к значительной потере данных, была применена стратегия замены unknown на самое частое значение (моду) в каждом столбце. Это позволило сохранить все наблюдения.
3. Коррекция типов данных:

```
Преобразование завершено. Финальная информация о датасете:
<class 'pandas.core.frame.DataFrame'>
Index: 41176 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41176 non-null  int64
 1   job             41176 non-null  category
 2   marital         41176 non-null  category
 3   education       41176 non-null  category
 4   default         41176 non-null  category
 5   housing         41176 non-null  category
 6   loan            41176 non-null  category
 7   contact         41176 non-null  category
 8   month           41176 non-null  category
 9   day_of_week     41176 non-null  category
 10  duration        41176 non-null  int64
 11  campaign        41176 non-null  int64
 12  pdays           41176 non-null  int64
 13  previous        41176 non-null  int64
 14  poutcome        41176 non-null  category
 15  emp.var.rate    41176 non-null  float64
 16  cons.price.idx  41176 non-null  float64
 17  cons.conf.idx   41176 non-null  float64
 18  euribor3m       41176 non-null  float64
 19  nr.employed     41176 non-null  float64
 20  y               41176 non-null  category
dtypes: category(11), float64(5), int64(5)
memory usage: 3.9 MB
```

    - Решение: Все столбцы типа object были преобразованы в category для оптимизации использования памяти (сокращение с 6.6+ MB до 3.9 MB) и семантической корректности.

Итоговый, очищенный датасет имеет размерность 41176 строк и 21 столбец.

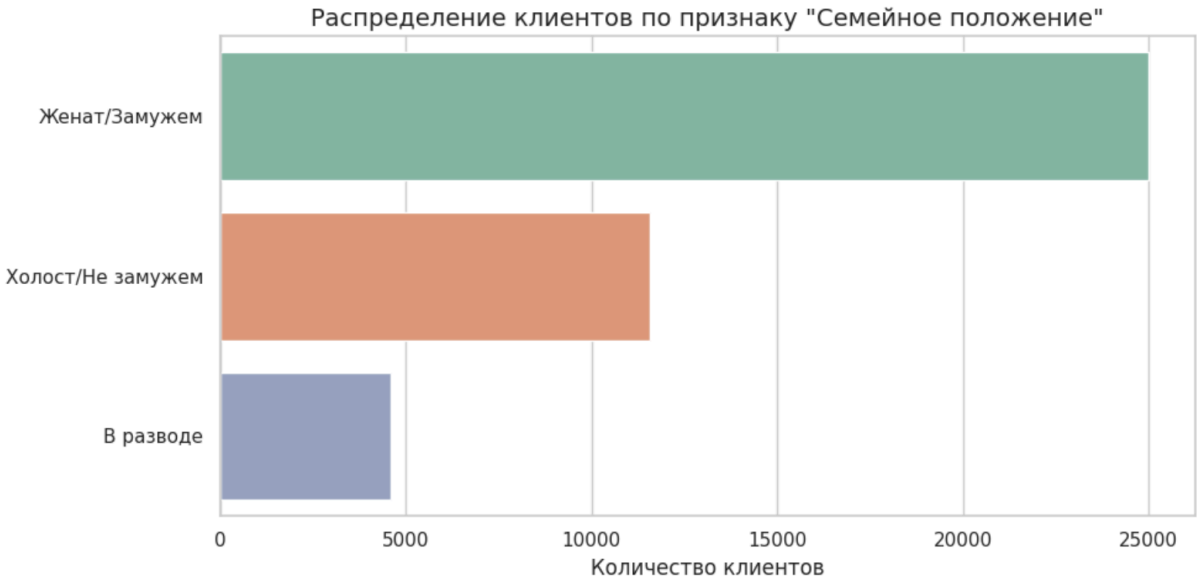
Этап 2: Однофакторный анализ (Анализ признаков по отдельности)

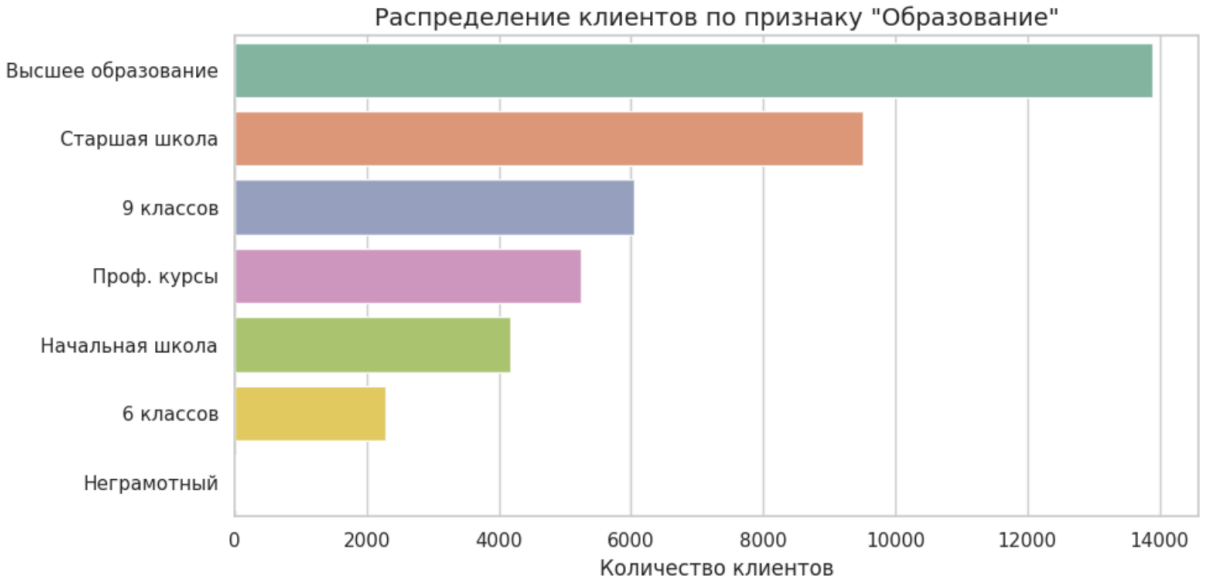
- Числовые признаки:

--- Численный анализ формы распределения ---

	Skewness	Kurtosis	Интерпретация асимметрии	Интерпретация куртозиса (хвостов)
Возраст	0.784560	0.791113	Умеренная правая (0.5-1)	Нормальные (-1-1)
Длительность звонка (сек)	3.262808	20.243771	Сильная правая (>1)	Очень тяжелые/длинные (>3)
Кол-во контактов в кампании	4.762044	36.971857	Сильная правая (>1)	Очень тяжелые/длинные (>3)
Дней с прошлого контакта	-4.921386	22.221553	Сильная левая (<-1)	Очень тяжелые/длинные (>3)
Кол-во прошлых контактов	3.831396	20.102164	Сильная правая (>1)	Очень тяжелые/длинные (>3)
Изм. уровня занятости	-0.724061	-1.062698	Умеренная левая (-1..-0.5)	Легкие/короткие (<-1)
Индекс потреб. цен	-0.230853	-0.829851	Почти симметрично (-0.5-0.5)	Нормальные (-1-1)
Индекс потреб. доверия	0.302876	-0.359097	Почти симметрично (-0.5-0.5)	Нормальные (-1-1)
Ставка Euribor 3M	-0.709194	-1.406791	Умеренная левая (-1..-0.5)	Легкие/короткие (<-1)
Число занятых	-1.044317	-0.003540	Сильная левая (<-1)	Нормальные (-1-1)

  - Анализ асимметрии и куртозиса показал, что поведенческие признаки duration (длительность), campaign (число контактов) и previous (прошлые контакты) имеют экстремально сильную правостороннюю асимметрию (коэффициент > 3) и очень "тяжелые" хвосты (куртозис > 20). Это свидетельствует о наличии большого количества выбросов.
  - Социально-экономические показатели, напротив, имеют распределения, близкие к симметричным или с "легкими" хвостами.
- Категориальные признаки:


Распределение клиентов по признаку "Семейное положение"
Женат/Замужем
Холост/Не замужем
В разводе
0
5000
10000
15000
20000
25000
Количество клиентов

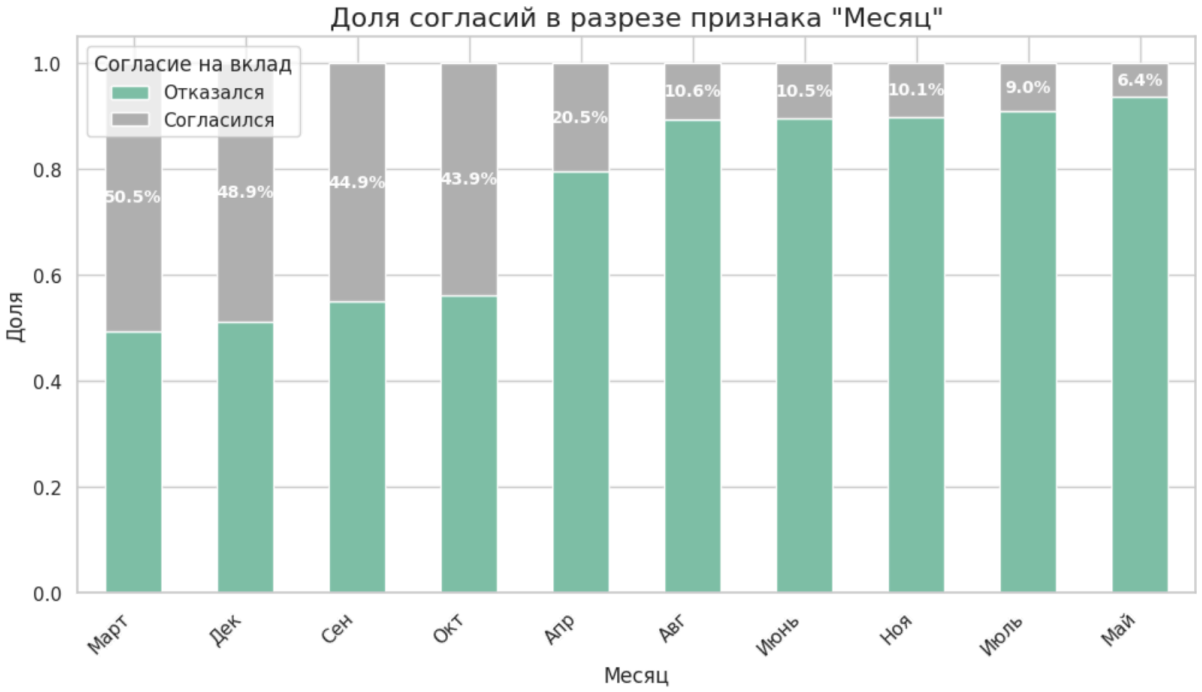

Распределение клиентов по признаку "Образование"
Высшее образование
Старшая школа
9 классов
Проф. курсы
Начальная школа
6 классов
Неграмотный
0
2000
4000
6000
8000
10000
12000
14000
Количество клиентов

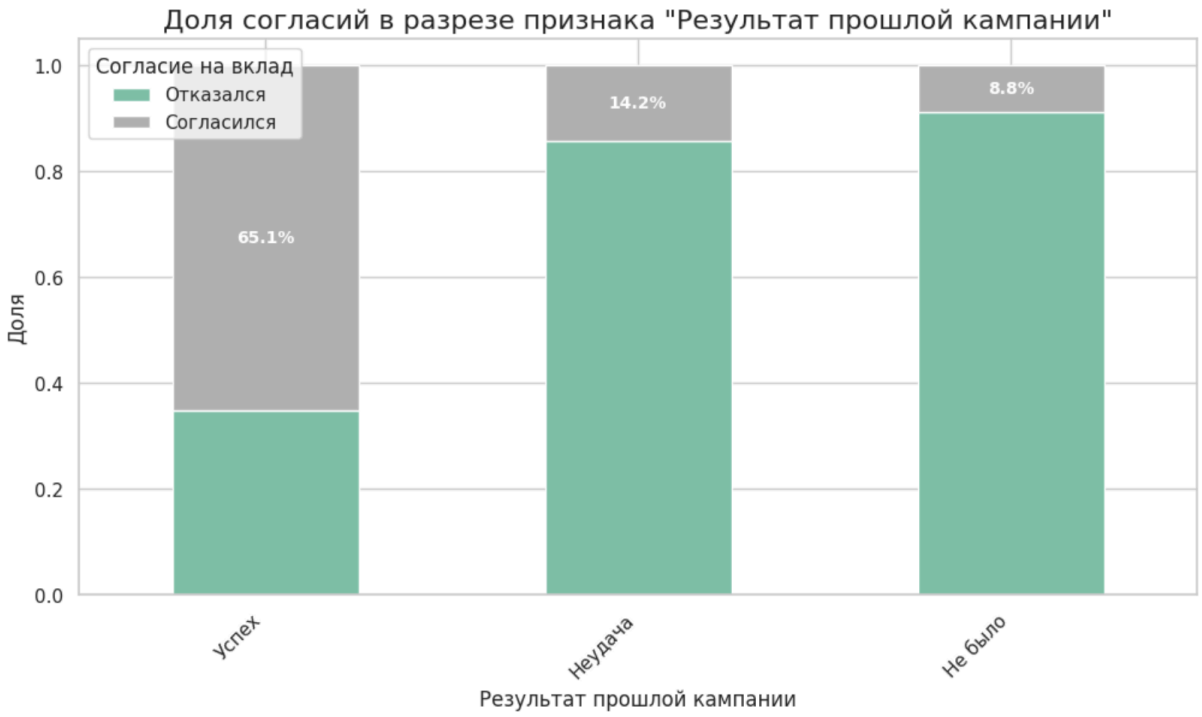

Распределение клиентов по признаку "Согласие на вклад"
no
yes
0
5000
10000
15000
20000
25000
30000
35000
Количество клиентов

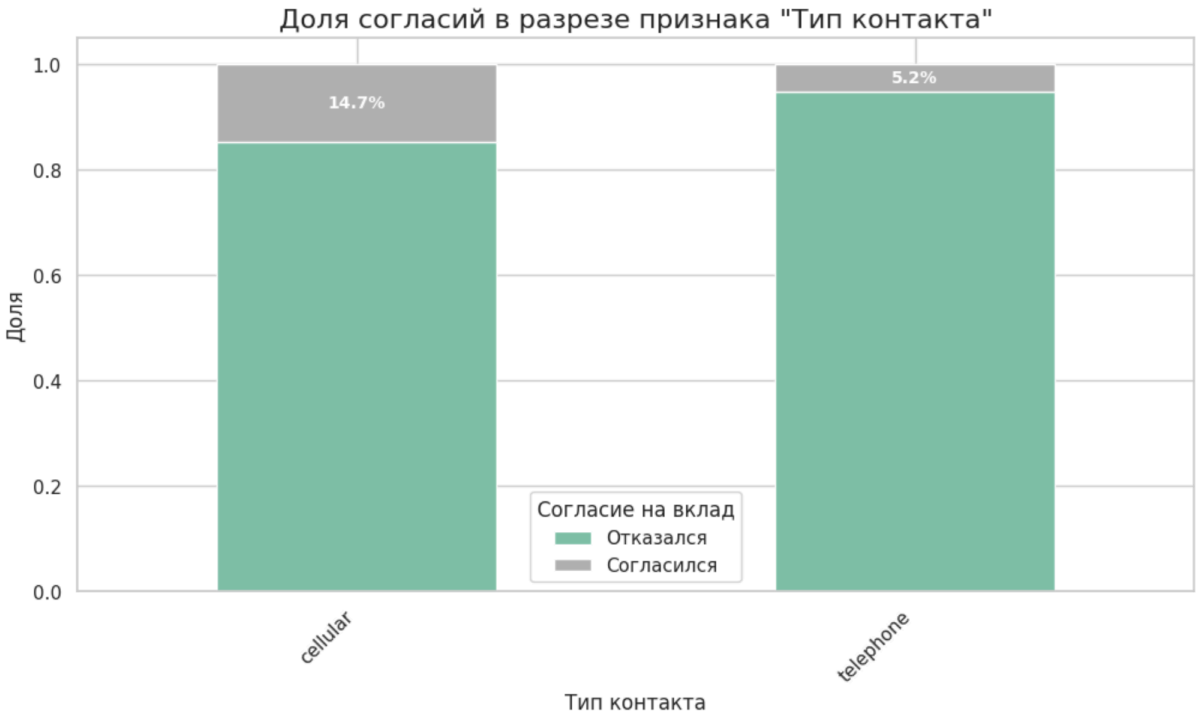
- Выявлен "портрет" среднестатистического клиента: административный работник (admin., 26.11%), состоящий в браке (married, 60.72%), с университетским образованием (university.degree, 33.74%).
- Ключевое наблюдение: Обнаружен сильный дисбаланс классов в целевой переменной : 88.73% клиентов ответили "нет" и только 11.27% — "да". Это является главной проблемой для будущего моделирования.

Этап 3: Двухфакторный анализ (Анализ взаимосвязей)

- Влияние на целевую переменную y:



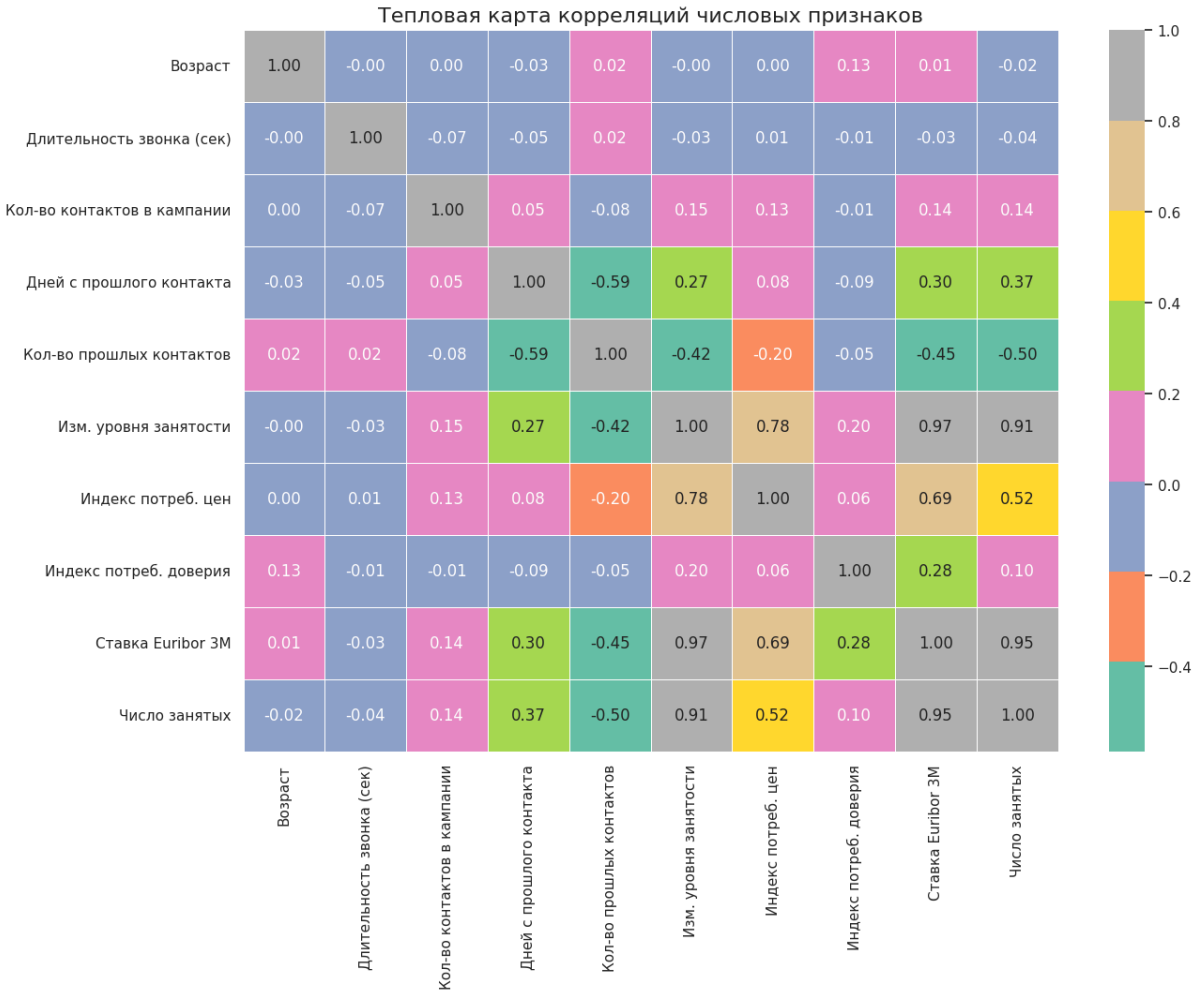





```
Результат теста 'duration' и 'campaign':

U-статистика: 1694100476.000
P-value: 0.000
Размеры выборок: 41,176 vs 41,176
Медиана 'duration': 180.00
Медиана 'campaign': 2.00
Среднее 'duration': 258.32
Среднее 'campaign': 2.57
Вывод: p-value < 0.05, распределения статистически значимо различаются.
Признаки 'duration' и 'campaign' имеют разные распределения

Результат теста 'duration' и 'previous':

U-статистика: 1695356733.000
P-value: 0.000
Размеры выборок: 41,176 vs 41,176
Медиана 'duration': 180.00
Медиана 'previous': 0.00
Среднее 'duration': 258.32
Среднее 'previous': 0.17
Вывод: p-value < 0.05, распределения статистически значимо различаются.
Признаки 'duration' и 'previous' имеют разные распределения

Результат теста 'duration' и 'pdays':

U-статистика: 100330959.500
P-value: 0.000
Размеры выборок: 41,176 vs 41,176
Медиана 'duration': 180.00
Медиана 'pdays': 999.00
Среднее 'duration': 258.32
Среднее 'pdays': 962.46
Вывод: p-value < 0.05, распределения статистически значимо различаются.
```

- Наиболее заинтересованными в банковских услугах являются студенты, пенсионеры и безработные. Это можепт быть связано с нехваткой денег.
- Одинокие люди пользуются банковскими услугами чаще. Возможно, они больше заинтересованны в построении своей карьеры и бизнеса. Другое предпололжение заключается в том, что обеспеченные люди чаще заводят семьи и имеют более стабильную жизнь.
- При звонке на мобильный телефон процент согласия выше почти в 3 раза

- Зависимость согласия от месяца, в который произошла связь с клиентом, довольно сильная и хаотичная
- Статистический U-тест Манна-Уитни показал, что различия в распределениях duration, campaign и previous для групп "yes" и "no" являются статистически значимыми ($p < 0.001$), а не случайными.



- Корреляции числовых признаков (Связь между внешними факторами (параметрами экономического контекста) не анализировалась, так как не связна с целью исследования - клиентами):
  - Признаки 'pdays' и 'previous имеют сильную отрицательную корреляцию (-0.59). Это логично, так как чем чаще клиенту звонили,
  - В остальном, числовые признаки почти не кореллируют, что предполагает их независимость.

Этап 4: Подготовка данных к моделированию

На данном этапе были формализованы стратегии преобразования данных для их подготовки к использованию в моделях машинного обучения.

- Стратегия работы с пропусками: Как было указано в Этапе 1, стратегия заключалась в замене "скрытых" пропусков (unknown) на моду, что является предпочтительным по сравнению с удалением строк.
- Стратегия кодирования категорий: Обоснован выбор метода One-Hot Encoding. Он преобразует каждую категорию в отдельный бинарный признак (0/1), что позволяет моделям корректно работать с номинальными данными без установления ложных порядковых зависимостей.
- Стратегия масштабирования числовых признаков: Обоснован выбор метода StandardScaler. Он приводит все числовые признаки к единому масштабу (среднее=0, ст. отклонение=1), что необходимо для корректной работы алгоритмов, чувствительных к масштабу признаков (например, Logistic Regression, SVM).

```
--- 8. Работа с пропусками ---
Стратегия: На начальном этапе анализа 'скрытые' пропуски (значения 'unknown')
были заменены на самое частое значение (моду) в соответствующем столбце.
Это позволило сохранить все строки данных, избежав потери информации.

--- 9. Кодирование категорий ---
Подход: Для преобразования категориальных признаков в числовой вид будет использован метод One-Hot Encoding.
Этот метод создает новые бинарные столбцы (0/1) для каждой категории, что позволяет избежать
неявного установления порядка между категориями (например, 'admin' не 'больше' чем 'student').
Это является стандартным и наиболее надежным подходом для большинства моделей машинного обучения.

--- 10. Масштабирование числовых признаков ---
Подход: Для числовых признаков будет применена стандартизация (StandardScaler).
Она приводит все признаки к единому масштабу (среднее=0, стандартное отклонение=1).
Это необходимо для корректной работы моделей, чувствительных к масштабу, таких как логистическая регрессия или SVM.

--- Демонстрация результата подготовки данных ---
Размерность данных после One-Hot Encoding и масштабирования: (41176, 57)
Итог: Получен полностью числовой, чистый датасет, готовый для моделирования.
Структура признаков сохранена в объекте 'preprocessor'.
```

1. ИСХОДНЫЕ ДАННЫЕ (первые 5 строк):

	Возраст	Профессия	Семейное положение	Образование	Кредитный дефолт	Ипотека	Потребительский кредит	Тип контакта	Месяц	День недели	Длительность звонка (сек)	Кол-во контактов в кампании	Дней с прошлого контакта	Кол-во прошлых контактов
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0
1	57	services	married	high.school	no	no	no	telephone	may	mon	149	1	999	0
2	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0
4	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0

2. ДАННЫЕ ПОСЛЕ ПРЕОБРАЗОВАНИЯ (первые 5 строк):
- Числовые столбцы (с префиксом 'num__') теперь стандартизированы (значения распределены вокруг 0).
- Категориальные столбцы (с префиксом 'cat__') превратились в множество бинарных колонок (0 или 1).
Например, 'job' превратился в 'cat__job_admin.', 'cat__job_blue-collar' и т.д.

	num__age	num__duration	num__campaign	num__pdays	num__previous	num__emp.var.rate	num__cons.price.idx	num__cons.conf.idx	num__euribor3m	num__nr.e
0	1.533143	0.010352	-0.565963	0.195443	-0.349551	0.648101	0.722628	0.886568	0.712463	
1	1.629107	-0.421577	-0.565963	0.195443	-0.349551	0.648101	0.722628	0.886568	0.712463	
2	-0.290177	-0.124626	-0.565963	0.195443	-0.349551	0.648101	0.722628	0.886568	0.712463	
3	-0.002284	-0.413864	-0.565963	0.195443	-0.349551	0.648101	0.722628	0.886568	0.712463	
4	1.533143	0.187751	-0.565963	0.195443	-0.349551	0.648101	0.722628	0.886568	0.712463	

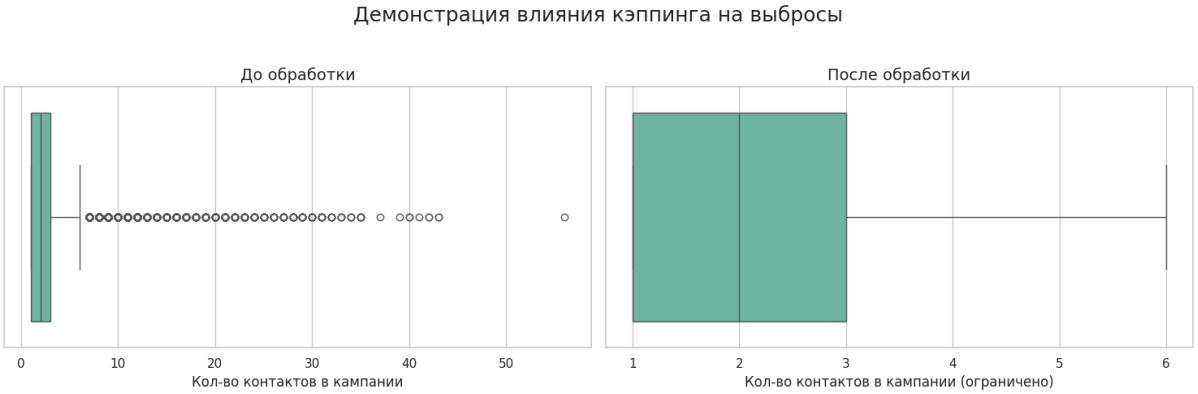
В результате применения этих стратегий исходный датасет преобразуется в полностью числовую матрицу размером (41176, 57), готовую к моделированию.

Этап 5: Анализ выбросов

```
Анализ age:
  Q1: 32.00, Q3: 47.00, IQR: 15.00
  Границы нормальных значений: от 9.50 до 69.50
  🤕 Найдено выбросов: 468 (1.14%)

Анализ duration:
  Q1: 102.00, Q3: 319.00, IQR: 217.00
  Границы нормальных значений: от -223.50 до 644.50
  😨 Найдено выбросов: 2963 (7.20%)

Анализ campaign:
  Q1: 1.00, Q3: 3.00, IQR: 2.00
  Границы нормальных значений: от -2.00 до 6.00
  😨 Найдено выбросов: 2406 (5.84%)

Анализ pdays:
  Q1: 999.00, Q3: 999.00, IQR: 0.00
  Границы нормальных значений: от 999.00 до 999.00
  🤕 Найдено выбросов: 1515 (3.68%)

Анализ previous:
  Q1: 0.00, Q3: 0.00, IQR: 0.00
  Границы нормальных значений: от 0.00 до 0.00
  😨 Найдено выбросов: 5625 (13.66%)

Анализ emp.var.rate:
  Q1: -1.80, Q3: 1.40, IQR: 3.20
  Границы нормальных значений: от -6.60 до 6.20
  🤕 Выбросов не найдено.

Анализ cons.price.idx:
  Q1: 93.08, Q3: 93.99, IQR: 0.92
  Границы нормальных значений: от 91.70 до 95.37
  🤕 Выбросов не найдено.

Анализ cons.conf.idx:
  Q1: -42.70, Q3: -36.40, IQR: 6.30
  Границы нормальных значений: от -52.15 до -26.95
  🤕 Найдено выбросов: 446 (1.08%)

Анализ euribor3m:
  Q1: 1.34, Q3: 4.96, IQR: 3.62
  Границы нормальных значений: от -4.08 до 10.39
  🤕 Выбросов не найдено.

Анализ nr.employed:
  Q1: 5099.10, Q3: 5228.10, IQR: 129.00
  Границы нормальных значений: от 4905.60 до 5421.60
  🤕 Выбросов не найдено.
```

- Обнаружение: Формальный анализ по методу межквартильного размаха (IQR) подтвердил наличие значительного числа выбросов. Наиболее проблемные признаки:
    - previous: 13.66% значений являются выбросами.

    - duration: 7.20% значений являются выбросами.
    - campaign: 5.84% значений являются выбросами.
- Стратегия обработки:
    - Обоснование: Удаление такого большого количества строк (более 10%) привело бы к существенной потере информации. Поэтому была рассмотрена альтернативная стратегия.
    - Решение: Была продемонстрирована стратегия "кэппинга" (ограничения). На примере признака campaign было показано, как замена всех значений, превышающих верхнюю границу (6.0), на это пороговое значение влияет на статистику:
        - До: mean=2.57, max=56.0
        - После: mean=2.28, max=6.0
    - Данный подход позволяет снизить влияние экстремальных значений, делая распределение более робастным, но при этом сохранить все наблюдения для дальнейшего анализа.



Этап 6: Итоговые выводы и гипотезы

Ключевые выводы:

1. Данные успешно очищены от дубликатов и скрытых пропусков (unknown).
2. Главная особенность датасета — сильный дисбаланс целевой переменной (89% "нет" / 11% "да").
3. Самым сильным предиктором успеха является результат предыдущей кампании (poutcome).
4. Студенты и пенсионеры — наиболее отзывчивые социальные группы.
5. Чрезмерная настойчивость (campaign) негативно влияет на конверсию.
6. Социально-экономические показатели сильно скоррелированы между собой (мультиколлинеарность).
7. Для признаков с большим количеством выбросов предложена стратегия кэппинга вместо удаления.

Гипотезы для моделирования:

1. Гипотеза 1: Выдвигается гипотеза, что согласие клиента можно спрогнозировать. Ожидается, что ключевыми факторами для предсказательной модели станут результат прошлой кампании (poutcome), длительность звонка (duration) и месяц контакта (month), поскольку в ходе разведочного анализа именно они продемонстрировали наибольшее влияние на итоговое решение.
2. Гипотеза 2: Признаки, описывающие историю и контекст взаимодействия клиента с банком, будут иметь значительно большую предсказательную силу, чем его статичные социально-демографические данные

Воспроизводимость:

- Для всех случайных процессов в дальнейшем будет использован random_state = 42.
- Версии ключевых библиотек:
    - Python: 3.12.11
    - pandas: 2.2.2
    - numpy: 2.0.2
    - seaborn: 0.13.2
    - scikit-learn: 1.6.1

```
--- Воспроизводимость ---
Для всех случайных процессов будет использован random_state = 42

--- Версии библиотек ---
Python: 3.12.11
pandas: 2.2.2
numpy: 2.0.2
seaborn: 0.13.2
scikit-learn: 1.6.1
```