German Credit Risk Dataset

Разведочный анализ данных (EDA)

Цель проекта:
Провести разведочный анализ данных (EDA), чтобы понять структуру датасета, выявить ключевые факторы кредитного риска и принять решения по предобработке данных перед построением моделей машинного обучения.

Целевая переменная:

- `credit_risk = 1` — надёжный (хороший) клиент
- `credit_risk = 0` — рискованный (плохой) клиент

In [ ]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

1. Обзор датасета

На данном этапе мы изучаем общую структуру данных:

- количество наблюдений
- типы признаков
- наличие пропусков

In [2]:
```
columns = [
    "checking_account",
    "duration_months",
    "credit_history",
    "purpose",
    "credit_amount",
    "savings_account",
    "employment_duration",
    "installment_rate",
    "personal_status",
    "other_debtors",
    "present_residence",
    "property",
    "age",
    "other_installment_plans",
    "housing",
    "existing_credits",
    "job",
    "num_dependents",
    "telephone",
    "foreign_worker",
    "credit_risk"
]
df= pd.read_csv(
```

```
    "C:\\Users\\nurs\\OneDrive\\Рабочий стол\\german.data",
    sep=" ",
    header=None,
    names=columns
)

df.head()
```

Out[2]:

	checking_account	duration_months	credit_history	purpose	credit_amount	savings_
0	A11	6	A34	A43	1169	
1	A12	48	A32	A43	5951	
2	A14	12	A34	A46	2096	
3	A11	42	A32	A42	7882	
4	A11	24	A33	A40	4870	

5 rows × 21 columns

In [3]:

```
# ================================
# 2. TARGET TRANSFORMATION
# 1 = good, 2 = bad  →  1 = good, 0 = bad
# ================================

df["credit_risk"] = df["credit_risk"].map({1: 1, 2: 0})

# ================================
# 3. FULL HUMAN-READABLE MAPPING
# ================================

df["checking_account"] = df["checking_account"].map({
    "A11": "< 0 DM",
    "A12": "0–200 DM",
    "A13": ">= 200 DM",
    "A14": "no checking account"
})

df["credit_history"] = df["credit_history"].map({
    "A30": "no credits / all paid",
    "A31": "all credits paid back",
    "A32": "existing credits paid",
    "A33": "delay in the past",
    "A34": "critical account"
})

df["purpose"] = df["purpose"].map({
    "A40": "car (new)",
    "A41": "car (used)",
    "A42": "furniture/equipment",
    "A43": "radio/TV",
    "A44": "domestic appliances",
    "A45": "repairs",
    "A46": "education",
    "A47": "vacation",
    "A48": "retraining",
    "A49": "business",
```

```
    "A410": "other"
})

df["savings_account"] = df["savings_account"].map({
    "A61": "< 100 DM",
    "A62": "100-500 DM",
    "A63": "500-1000 DM",
    "A64": ">= 1000 DM",
    "A65": "unknown / none"
})

df["employment_duration"] = df["employment_duration"].map({
    "A71": "unemployed",
    "A72": "< 1 year",
    "A73": "1-4 years",
    "A74": "4-7 years",
    "A75": ">= 7 years"
})

df["personal_status"] = df["personal_status"].map({
    "A91": "male divorced/separated",
    "A92": "female divorced/separated/married",
    "A93": "male single",
    "A94": "male married/widowed",
    "A95": "female single"
})

df["other_debtors"] = df["other_debtors"].map({
    "A101": "none",
    "A102": "co-applicant",
    "A103": "guarantor"
})

df["property"] = df["property"].map({
    "A121": "real estate",
    "A122": "life insurance / savings",
    "A123": "car / other",
    "A124": "unknown / none"
})

df["other_installment_plans"] = df["other_installment_plans"].map({
    "A141": "bank",
    "A142": "stores",
    "A143": "none"
})

df["housing"] = df["housing"].map({
    "A151": "rent",
    "A152": "own",
    "A153": "for free"
})

df["job"] = df["job"].map({
    "A171": "unemployed / unskilled (non-resident)",
    "A172": "unskilled (resident)",
    "A173": "skilled employee",
    "A174": "management / highly qualified"
})

df["telephone"] = df["telephone"].map({
```

```
    "A191": "no",
    "A192": "yes"
})

df["foreign_worker"] = df["foreign_worker"].map({
    "A201": "yes",
    "A202": "no"
})

# ==============================
# 4. FINAL CHECK
# ==============================

print(df.head())
```

```
     checking_account  duration_months         credit_history  \
0              < 0 DM                6       critical account
1            0–200 DM               48  existing credits paid
2  no checking account               12       critical account
3              < 0 DM               42  existing credits paid
4              < 0 DM               24      delay in the past

              purpose  credit_amount savings_account employment_duration  \
0            radio/TV           1169  unknown / none          >= 7 years
1            radio/TV           5951        < 100 DM           1–4 years
2           education           2096        < 100 DM           4–7 years
3  furniture/equipment           7882        < 100 DM           4–7 years
4           car (new)            4870        < 100 DM           1–4 years

   installment_rate                   personal_status other_debtors  ...  \
0                 4                       male single          none  ...
1                 2  female divorced/separated/married          none  ...
2                 2                       male single          none  ...
3                 2                       male single     guarantor  ...
4                 3                       male single          none  ...

                  property age  other_installment_plans   housing  \
0              real estate  67                     none       own
1              real estate  22                     none       own
2              real estate  49                     none       own
3  life insurance / savings  45                     none  for free
4           unknown / none  53                     none  for free

  existing_credits                  job num_dependents  telephone  \
0                2      skilled employee              1        yes
1                1      skilled employee              1         no
2                1  unskilled (resident)              2         no
3                1      skilled employee              2         no
4                2      skilled employee              2         no

  foreign_worker credit_risk
0            yes           1
1            yes           0
2            yes           1
3            yes           1
4            yes           0

[5 rows x 21 columns]
```

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 21 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   checking_account         1000 non-null   object
 1   duration_months          1000 non-null   int64 
 2   credit_history           1000 non-null   object
 3   purpose                  1000 non-null   object
 4   credit_amount            1000 non-null   int64 
 5   savings_account          1000 non-null   object
 6   employment_duration      1000 non-null   object
 7   installment_rate         1000 non-null   int64 
 8   personal_status          1000 non-null   object
 9   other_debtors            1000 non-null   object
 10  present_residence        1000 non-null   int64 
 11  property                 1000 non-null   object
 12  age                      1000 non-null   int64 
 13  other_installment_plans  1000 non-null   object
 14  housing                  1000 non-null   object
 15  existing_credits         1000 non-null   int64 
 16  job                      1000 non-null   object
 17  num_dependents           1000 non-null   int64 
 18  telephone                1000 non-null   object
 19  foreign_worker           1000 non-null   object
 20  credit_risk              1000 non-null   int64 
dtypes: int64(8), object(13)
memory usage: 164.2+ KB
```

Выводы

- Датасет содержит **1000 наблюдений и 21 признак**
- Пропущенные значения отсутствуют
- В данных присутствуют:
  - числовые признаки (сумма кредита, срок, возраст и т.д.)
  - категориальные признаки (статус счёта, кредитная история, жильё и т.д.)

Данные являются чистыми и готовы к анализу.

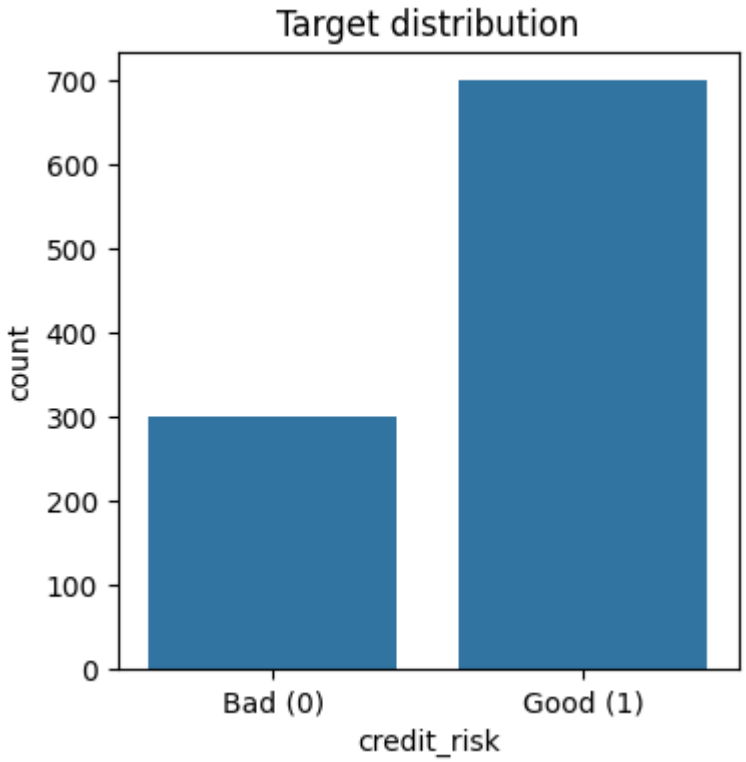
2. Анализ целевой переменной

Цель данного шага — понять распределение целевой переменной `credit_risk`.

In [5]:
```
target_counts = df["credit_risk"].value_counts()
target_ratio = df["credit_risk"].value_counts(normalize=True)
display(target_counts, target_ratio)
plt.figure(figsize=(4,4))
sns.countplot(x="credit_risk", data=df)
plt.xticks([0, 1], ["Bad (0)", "Good (1)"])
plt.title("Target distribution")
plt.show()
```

```
credit_risk
1    700
0    300
Name: count, dtype: int64
```

```
credit_risk
1    0.7
0    0.3
Name: proportion, dtype: float64
```



Выводы

- 70% клиентов являются надёжными
- 30% клиентов — рискованные
- Наблюдается **умеренный дисбаланс классов**

Это означает, что в дальнейшем:

- accuracy не является основной метрикой
- необходимо использовать ROC-AUC, PR-AUC и F1-score

Распределение Признаков

Здесь мы распределяли признаки на свои категории:

- Наш таргет
- Числовые
- Категориальные

In [6]:
```
target_col = "credit_risk"
num_cols = df.select_dtypes(include="int64").columns.drop(target_col).tolist()
cat_cols = df.select_dtypes(include="object").columns.tolist()
num_cols, cat_cols
```

Out[6]:
```
(['duration_months',
  'credit_amount',
  'installment_rate',
  'present_residence',
  'age',
  'existing_credits',
  'num_dependents'],
 ['checking_account',
  'credit_history',
  'purpose',
  'savings_account',
  'employment_duration',
  'personal_status',
  'other_debtors',
  'property',
  'other_installment_plans',
  'housing',
  'job',
  'telephone',
  'foreign_worker'])
```

In [7]:
```
df[cat_cols].nunique().sort_values(ascending=False)
```

Out[7]:
```
purpose                   10
savings_account            5
credit_history             5
employment_duration        5
checking_account           4
personal_status            4
property                   4
job                        4
other_debtors              3
other_installment_plans    3
housing                    3
telephone                  2
foreign_worker             2
dtype: int64
```

3. Анализ числовых признаков

На данном этапе анализируются числовые признаки с целью:

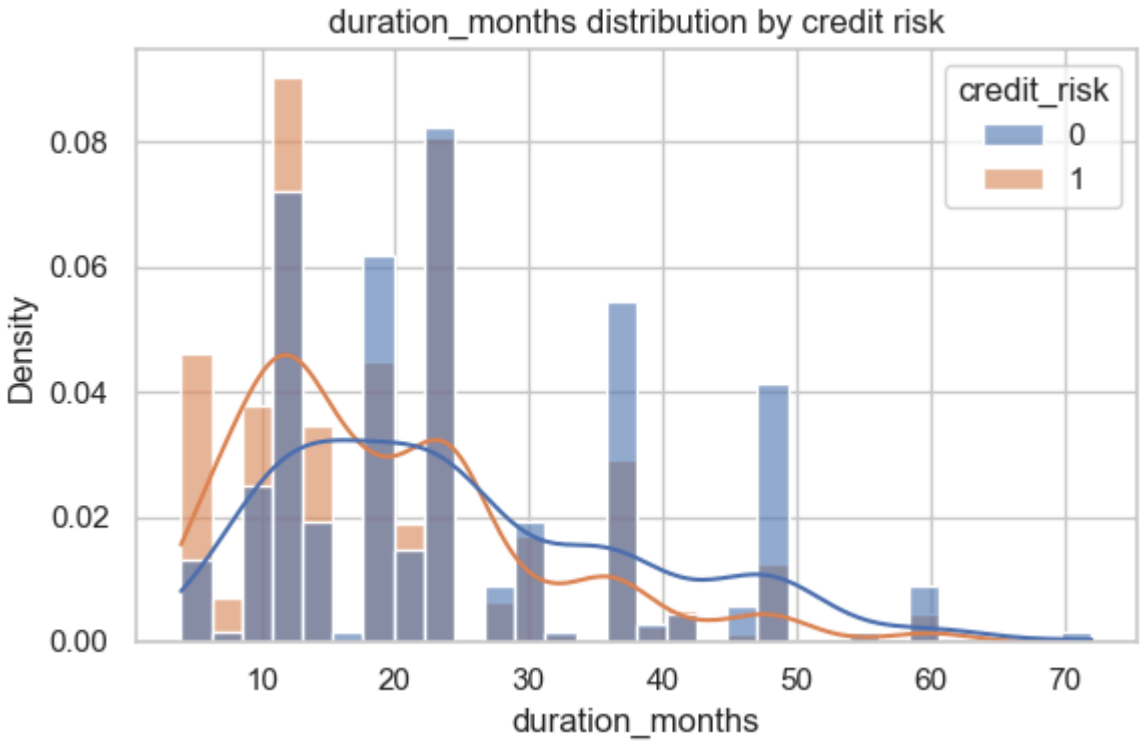
- изучить распределения
- выявить асимметрию и выбросы
- оценить связь с кредитным риском

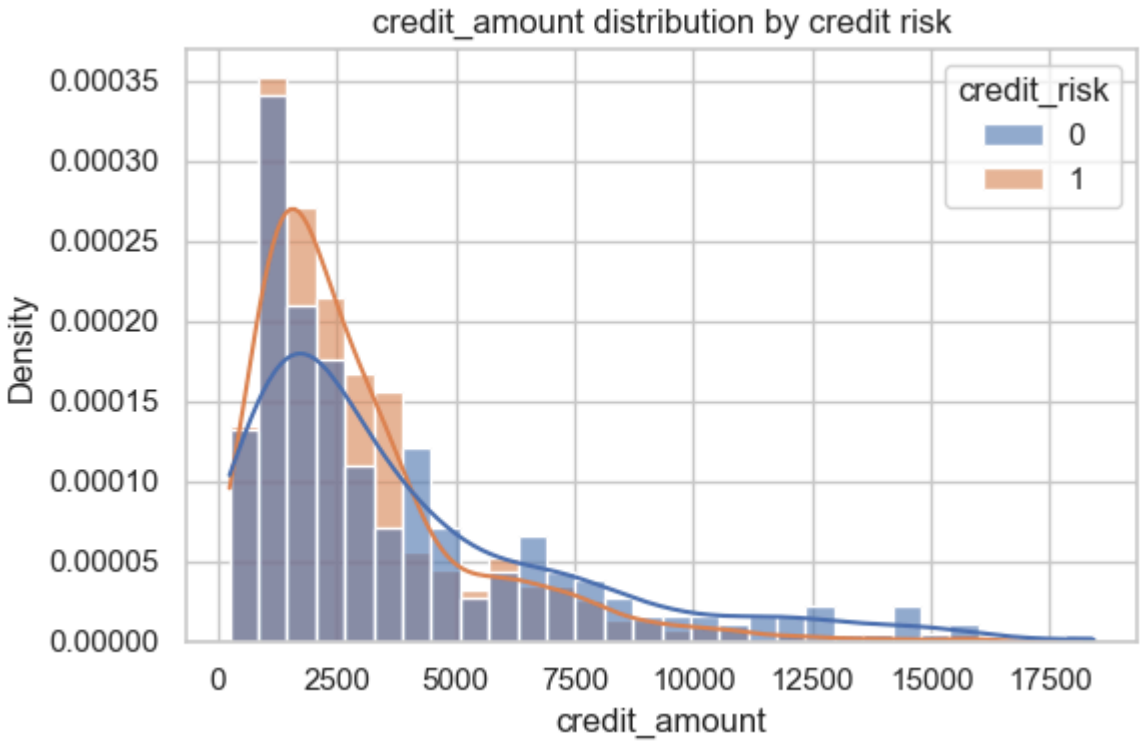
In [11]:
```
df.describe()
```

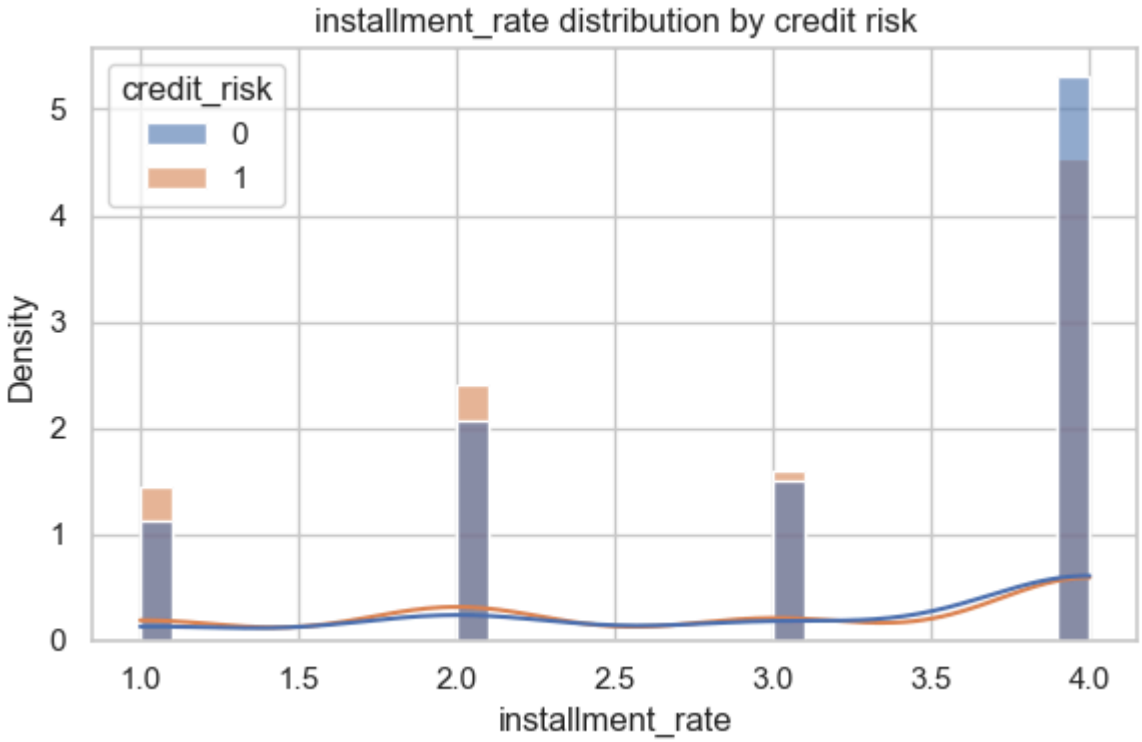
Out[11]:

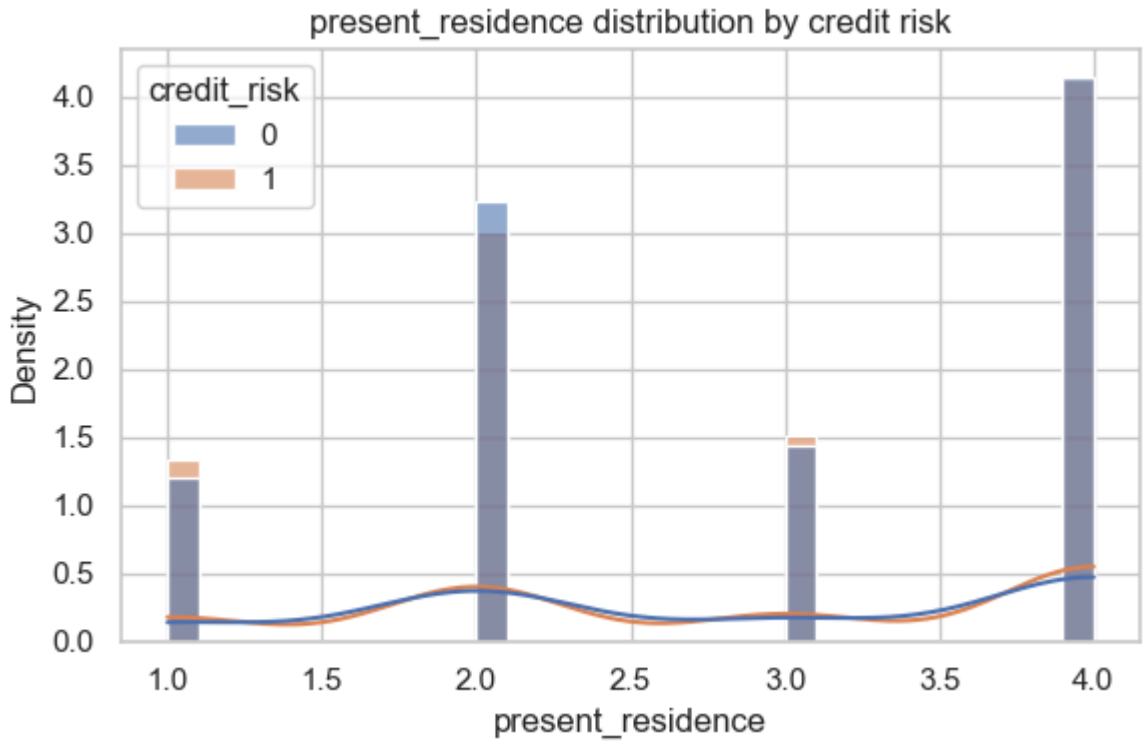
	duration_months	credit_amount	installment_rate	present_residence	age
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	20.903000	3271.258000	2.973000	2.845000	35.546000
std	12.058814	2822.736876	1.118715	1.103718	11.375469
min	4.000000	250.000000	1.000000	1.000000	19.000000
25%	12.000000	1365.500000	2.000000	2.000000	27.000000
50%	18.000000	2319.500000	3.000000	3.000000	33.000000
75%	24.000000	3972.250000	4.000000	4.000000	42.000000
max	72.000000	18424.000000	4.000000	4.000000	75.000000

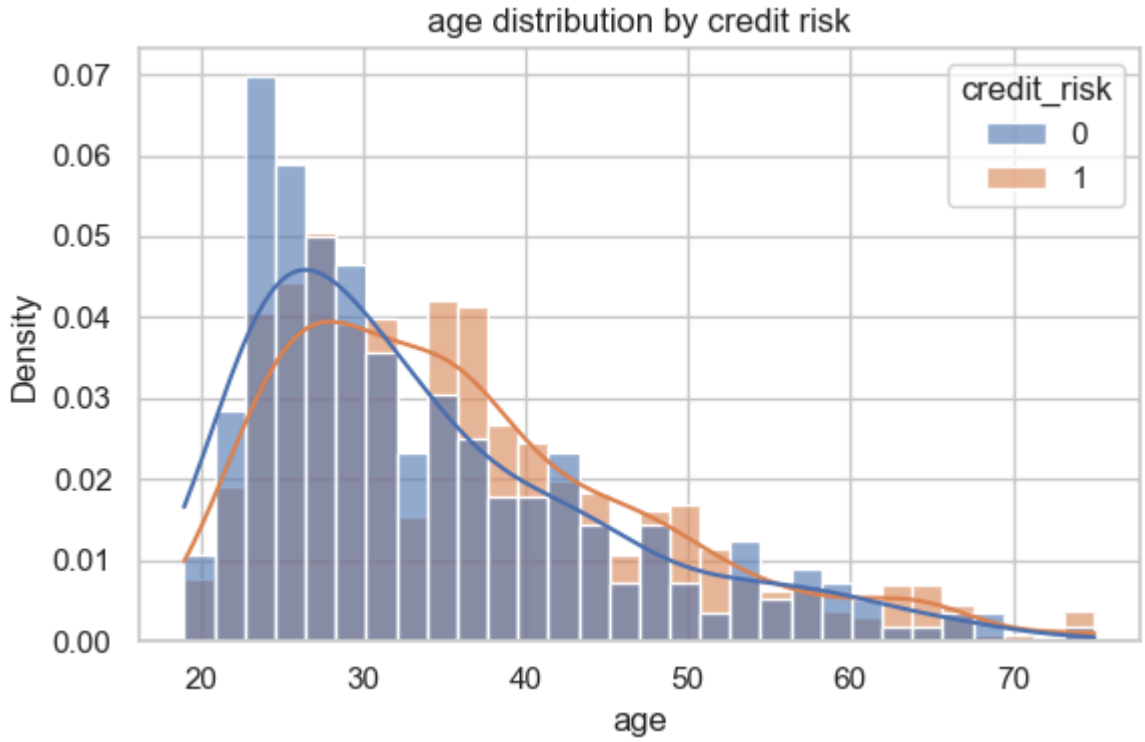
In [10]:
```
for col in num_cols:
    plt.figure(figsize=(6, 4))
    sns.histplot(
        data=df,
        x=col,
        hue="credit_risk",
        bins=30,
        kde=True,
        stat="density",
        common_norm=False,
        alpha=0.6
    )
    plt.title(f"{col} distribution by credit risk")
    plt.xlabel(col)
    plt.ylabel("Density")
    plt.tight_layout()
    plt.show()
```

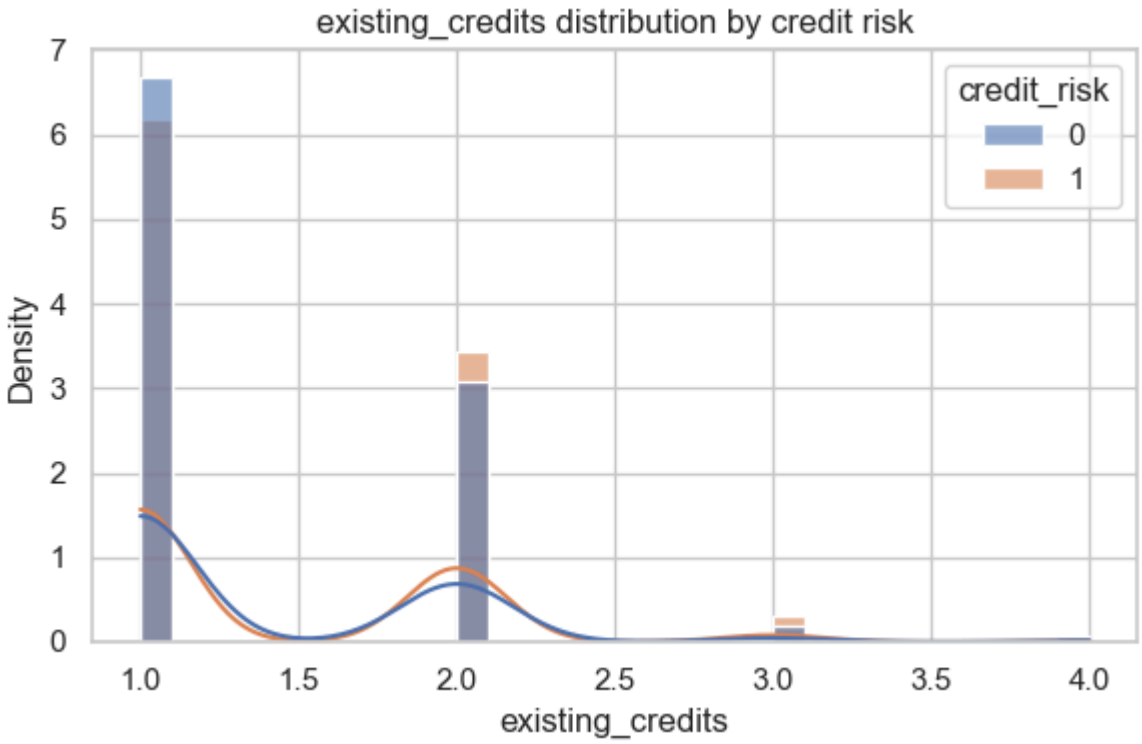


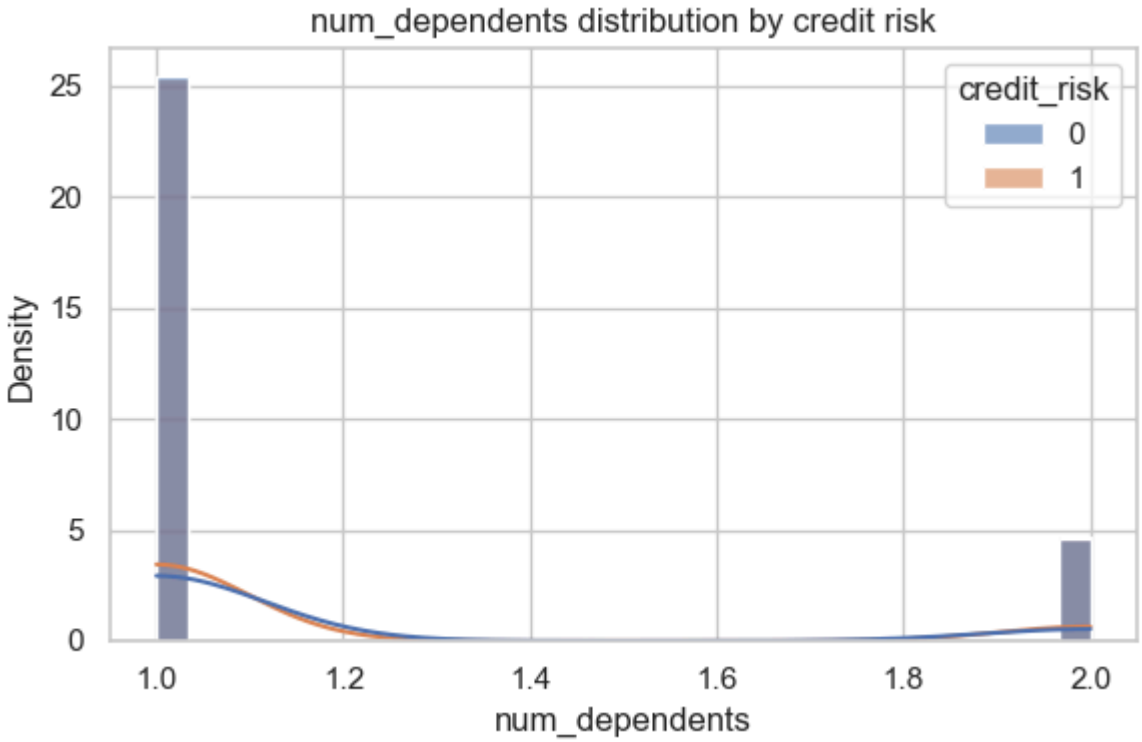

credit_amount distribution by credit risk
credit_risk
0
1
Density
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
2500
5000
7500
10000
12500
15000
17500
credit_amount


installment_rate distribution by credit risk
credit_risk
0
1
Density
5
4
3
2
1
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
installment_rate


present_residence distribution by credit risk
credit_risk
0
1
Density
present_residence
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0


age distribution by credit risk
credit_risk
0
1
Density
age
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
70





Итоги по числовым признакам

Числовые признаки демонстрируют различную степень асимметрии распределений. Наиболее выраженная правосторонняя асимметрия наблюдается у суммы кредита, что указывает на наличие редких, но крупных значений. Анализ показал, что увеличение суммы кредита и срока кредитования связано с ростом кредитного риска. Более молодые клиенты в среднем характеризуются более высоким уровнем дефолтов, тогда как возраст и другие показатели стабильности снижают риск. В совокупности числовые признаки хорошо отражают финансовую нагрузку клиента и являются информативными для модели.

4. Анализ категориальных признаков

Категориальные признаки анализируются путём оценки доли плохих и хороших клиентов в каждой категории.

In [ ]:
```
def cat_risk_analysis(df, col, target="credit_risk"):
    summary = (
        df.groupby(col)[target]
        .agg(["count", "mean"])
        .rename(columns={"mean": "good_rate"})
        .sort_values("good_rate")
    )
    summary["bad_rate"] = 1 - summary["good_rate"]
    return summary
```

In [13]:
```
cat_risk_analysis(df, "checking_account")
```

Out[13]:

checking_account	count	good_rate	bad_rate
< 0 DM	274	0.507299	0.492701
0–200 DM	269	0.609665	0.390335
>= 200 DM	63	0.777778	0.222222
no checking account	394	0.883249	0.116751

In [14]:
```
cat_risk_analysis(df, "credit_history")
```

Out[14]:

credit_history	count	good_rate	bad_rate
no credits / all paid	40	0.375000	0.625000
all credits paid back	49	0.428571	0.571429
existing credits paid	530	0.681132	0.318868
delay in the past	88	0.681818	0.318182
critical account	293	0.829352	0.170648

In [15]:
```
cat_risk_analysis(df, "purpose")
```

Out[15]:

purpose	count	good_rate	bad_rate
education	50	0.560000	0.440000
other	12	0.583333	0.416667
car (new)	234	0.619658	0.380342
repairs	22	0.636364	0.363636
business	97	0.649485	0.350515
domestic appliances	12	0.666667	0.333333
furniture/equipment	181	0.679558	0.320442
radio/TV	280	0.778571	0.221429
car (used)	103	0.834951	0.165049
retraining	9	0.888889	0.111111

In [17]:
```
cat_risk_analysis(df, "savings_account")
```

Out[17]:

savings_account	count	good_rate	bad_rate
< 100 DM	603	0.640133	0.359867
100–500 DM	103	0.669903	0.330097
unknown / none	183	0.825137	0.174863
500–1000 DM	63	0.825397	0.174603
>= 1000 DM	48	0.875000	0.125000

In [18]:
```
cat_risk_analysis(df, "employment_duration")
```

Out[18]:

employment_duration	count	good_rate	bad_rate
< 1 year	172	0.593023	0.406977
unemployed	62	0.629032	0.370968
1–4 years	339	0.693215	0.306785
>= 7 years	253	0.747036	0.252964
4–7 years	174	0.775862	0.224138

In [19]:
```
cat_risk_analysis(df, "personal_status")
```

Out[19]:

personal_status	count	good_rate	bad_rate
male divorced/separated	50	0.600000	0.400000
female divorced/separated/married	310	0.648387	0.351613
male married/widowed	92	0.728261	0.271739
male single	548	0.733577	0.266423

In [20]:
```
cat_risk_analysis(df, "property")
```

Out[20]:

property	count	good_rate	bad_rate
unknown / none	154	0.564935	0.435065
car / other	332	0.692771	0.307229
life insurance / savings	232	0.693966	0.306034
real estate	282	0.787234	0.212766

In [21]:
```
cat_risk_analysis(df, "housing")
```

Out[21]:

housing	count	good_rate	bad_rate
for free	108	0.592593	0.407407
rent	179	0.608939	0.391061
own	713	0.739130	0.260870

In [22]:
```
cat_risk_analysis(df, "job")
```

Out[22]:

job	count	good_rate	bad_rate
management / highly qualified	148	0.655405	0.344595
unemployed / unskilled (non-resident)	22	0.681818	0.318182
skilled employee	630	0.704762	0.295238
unskilled (resident)	200	0.720000	0.280000

In [23]:
```
cat_risk_analysis(df, "telephone")
```

Out[23]:

telephone	count	good_rate	bad_rate
no	596	0.686242	0.313758
yes	404	0.720297	0.279703

In [24]:
```
cat_risk_analysis(df, "foreign_worker")
```

Out[24]:

foreign_worker	count	good_rate	bad_rate
yes	963	0.692627	0.307373
no	37	0.891892	0.108108

Итоги по категориальным признакам

Категориальные признаки оказались одними из наиболее значимых факторов кредитного риска. Наиболее низкий риск наблюдается у клиентов с устойчивым финансовым положением: положительный баланс по счетам, наличие сбережений, длительный стаж работы, собственное жильё и имущество. Напротив, признаки финансовой нестабильности — отрицательный баланс, короткий трудовой стаж, отсутствие имущества или аренда жилья — связаны с повышенной вероятностью дефолта. Некоторые категориальные признаки демонстрируют более слабую связь с целевой переменной, однако в совокупности они могут улучшать качество модели.

Общий вывод

Проведённый EDA показал, что кредитный риск формируется сочетанием финансовой нагрузки, кредитной истории и социальной стабильности клиента. Датасет логичен, информативен и соответствует реальным банковским сценариям. Полученные выводы позволяют обоснованно перейти к этапу предобработки данных и построению моделей машинного обучения.