Predicting Company Bankruptcy Using Financial Data

In [62]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
import matplotlib.pyplot as plt
from scipy.stats import chi2_contingency
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, precision_score, recall_score, accuracy_score, f1_score
```

In [63]:
```
df=pd.read_csv('data.csv')
```

In [64]:
```
1 df
```

Out[64]:

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue	...	Net Income to Total Assets	a to
0	1	0.370594	0.424389	0.405750	0.601457	0.601457	0.998969	0.796887	0.808809	0.302646	...	0.716845	0.0(
1	1	0.464291	0.538214	0.516730	0.610235	0.610235	0.998946	0.797380	0.809301	0.303556	...	0.795297	0.0(
2	1	0.426071	0.499019	0.472295	0.601450	0.601364	0.998857	0.796403	0.808388	0.302035	...	0.774670	0.0⁴
3	1	0.399844	0.451265	0.457733	0.583541	0.583541	0.998700	0.796967	0.808966	0.303350	...	0.739555	0.0(
4	1	0.465022	0.538432	0.522298	0.598783	0.598783	0.998973	0.797366	0.809304	0.303475	...	0.795016	0.0(
...	...	...	...	...	...	...	...	...	...	...	...	...	
6814	0	0.493687	0.539468	0.543230	0.604455	0.604462	0.998992	0.797409	0.809331	0.303510	...	0.799927	0.0(
6815	0	0.475162	0.538269	0.524172	0.598308	0.598308	0.998992	0.797414	0.809327	0.303520	...	0.799748	0.0(
6816	0	0.472725	0.533744	0.520638	0.610444	0.610213	0.998984	0.797401	0.809317	0.303512	...	0.797778	0.0(
6817	0	0.506264	0.559911	0.554045	0.607850	0.607850	0.999074	0.797500	0.809399	0.303498	...	0.811808	0.0(
6818	0	0.493053	0.570105	0.549548	0.627409	0.627409	0.998080	0.801987	0.813800	0.313415	...	0.815956	0.0(

6819 rows × 96 columns

In [4]:
```
1 df.head()
```

Out[4]:

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue	...	Net Income to Total Assets	Tot asse to GN pri
0	1	0.370594	0.424389	0.405750	0.601457	0.601457	0.998969	0.796887	0.808809	0.302646	...	0.716845	0.0092
1	1	0.464291	0.538214	0.516730	0.610235	0.610235	0.998946	0.797380	0.809301	0.303556	...	0.795297	0.0083
2	1	0.426071	0.499019	0.472295	0.601450	0.601364	0.998857	0.796403	0.808388	0.302035	...	0.774670	0.0400
3	1	0.399844	0.451265	0.457733	0.583541	0.583541	0.998700	0.796967	0.808966	0.303350	...	0.739555	0.0032
4	1	0.465022	0.538432	0.522298	0.598783	0.598783	0.998973	0.797366	0.809304	0.303475	...	0.795016	0.0038

5 rows × 96 columns

In [5]:
```
1 df.tail()
```

Out[5]:

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue	...	Net Income to Total Assets	a to
6814	0	0.493687	0.539468	0.543230	0.604455	0.604462	0.998992	0.797409	0.809331	0.303510	...	0.799927	0.0
6815	0	0.475162	0.538269	0.524172	0.598308	0.598308	0.998992	0.797414	0.809327	0.303520	...	0.799748	0.0
6816	0	0.472725	0.533744	0.520638	0.610444	0.610213	0.998984	0.797401	0.809317	0.303512	...	0.797778	0.0
6817	0	0.506264	0.559911	0.554045	0.607850	0.607850	0.999074	0.797500	0.809399	0.303498	...	0.811808	0.0
6818	0	0.493053	0.570105	0.549548	0.627409	0.627409	0.998080	0.801987	0.813800	0.313415	...	0.815956	0.0

5 rows × 96 columns

In [6]:
```
1 df.shape
```

Out[6]: (6819, 96)

In [7]:
```
1 df.columns
```

Out[7]:
```
Index(['Bankrupt?', 'ROA(C) before interest and depreciation before interest',
       'ROA(A) before interest and % after tax',
       'ROA(B) before interest and depreciation after tax',
       'Operating Gross Margin', 'Realized Sales Gross Margin',
       'Operating Profit Rate', 'Pre-tax net Interest Rate',
       'After-tax net Interest Rate',
       'Non-industry income and expenditure/revenue',
       'Continuous interest rate (after tax)', 'Operating Expense Rate',
       'Research and development expense rate', 'Cash flow rate',
       'Interest-bearing debt interest rate', 'Tax rate (A)',
       'Net Value Per Share (B)', 'Net Value Per Share (A)',
       'Net Value Per Share (C)', 'Persistent EPS in the Last Four Seasons',
       'Cash Flow Per Share', 'Revenue Per Share (Yuan ¥)',
       'Operating Profit Per Share (Yuan ¥)',
       'Per Share Net profit before tax (Yuan ¥)',
       'Realized Sales Gross Profit Growth Rate',
       'Operating Profit Growth Rate', 'After-tax Net Profit Growth Rate',
       'Regular Net Profit Growth Rate', 'Continuous Net Profit Growth Rate',
       'Total Asset Growth Rate', 'Net Value Growth Rate',
       'Total Asset Return Growth Rate Ratio', 'Cash Reinvestment %',
       'Current Ratio', 'Quick Ratio', 'Interest Expense Ratio',
       'Total debt/Total net worth', 'Debt ratio %', 'Net worth/Assets',
       'Long-term fund suitability ratio (A)', 'Borrowing dependency',
       'Contingent liabilities/Net worth', 'Operating profit/Paid-in capital',
       'Net profit before tax/Paid-in capital',
       'Inventory and accounts receivable/Net value', 'Total Asset Turnover',
       'Accounts Receivable Turnover', 'Average Collection Days',
       'Inventory Turnover Rate (times)', 'Fixed Assets Turnover Frequency',
       'Net Worth Turnover Rate (times)', 'Revenue per person',
       'Operating profit per person', 'Allocation rate per person',
       'Working Capital to Total Assets', 'Quick Assets/Total Assets',
       'Current Assets/Total Assets', 'Cash/Total Assets',
       'Quick Assets/Current Liability', 'Cash/Current Liability',
       'Current Liability to Assets', 'Operating Funds to Liability',
       'Inventory/Working Capital', 'Inventory/Current Liability',
       'Current Liabilities/Liability', 'Working Capital/Equity',
       'Current Liabilities/Equity', 'Long-term Liability to Current Assets',
       'Retained Earnings to Total Assets', 'Total income/Total expense',
       'Total expense/Assets', 'Current Asset Turnover Rate',
       'Quick Asset Turnover Rate', 'Working capitcal Turnover Rate',
       'Cash Turnover Rate', 'Cash Flow to Sales', 'Fixed Assets to Assets',
```

```
       'Current Liability to Liability', 'Current Liability to Equity',
       'Equity to Long-term Liability', 'Cash Flow to Total Assets',
       'Cash Flow to Liability', 'CFO to Assets', 'Cash Flow to Equity',
       'Current Liability to Current Assets', 'Liability-Assets Flag',
       'Net Income to Total Assets', 'Total assets to GNP price',
       'No-credit Interval', 'Gross Profit to Sales',
       'Net Income to Stockholder's Equity', 'Liability to Equity',
       'Degree of Financial Leverage (DFL)',
       'Interest Coverage Ratio (Interest expense to EBIT)', 'Net Income Flag',
       'Equity to Liability'],
      dtype='object')
```

In [8]:
```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6819 entries, 0 to 6818
Data columns (total 96 columns):
 #   Column                                                  Non-Null Count  Dtype  
---  ------                                                  --------------  -----  
 0   Bankrupt?                                               6819 non-null   int64  
 1   ROA(C) before interest and depreciation before interest  6819 non-null   float64
 2   ROA(A) before interest and % after tax                  6819 non-null   float64
 3   ROA(B) before interest and depreciation after tax       6819 non-null   float64
 4   Operating Gross Margin                                  6819 non-null   float64
 5   Realized Sales Gross Margin                             6819 non-null   float64
 6   Operating Profit Rate                                   6819 non-null   float64
 7   Pre-tax net Interest Rate                               6819 non-null   float64
 8   After-tax net Interest Rate                             6819 non-null   float64
 9   Non-industry income and expenditure/revenue             6819 non-null   float64
 10  Continuous interest rate (after tax)                    6819 non-null   float64
 11  Operating Expense Rate                                  6819 non-null   float64
 12  Research and development expense rate                   6819 non-null   float64
 13  Cash flow rate                                          6819 non-null   float64
 14  Interest-bearing debt interest rate                     6819 non-null   float64
 15  Tax rate (A)                                            6819 non-null   float64
 16  Net Value Per Share (B)                                 6819 non-null   float64
 17  Net Value Per Share (A)                                 6819 non-null   float64
 18  Net Value Per Share (C)                                 6819 non-null   float64
 19  Persistent EPS in the Last Four Seasons                 6819 non-null   float64
 20  Cash Flow Per Share                                     6819 non-null   float64
 21  Revenue Per Share (Yuan ¥)                              6819 non-null   float64
 22  Operating Profit Per Share (Yuan ¥)                     6819 non-null   float64
 23  Per Share Net profit before tax (Yuan ¥)                6819 non-null   float64
 24  Realized Sales Gross Profit Growth Rate                 6819 non-null   float64
 25  Operating Profit Growth Rate                            6819 non-null   float64
 26  After-tax Net Profit Growth Rate                        6819 non-null   float64
 27  Regular Net Profit Growth Rate                          6819 non-null   float64
 28  Continuous Net Profit Growth Rate                       6819 non-null   float64
 29  Total Asset Growth Rate                                 6819 non-null   float64
 30  Net Value Growth Rate                                   6819 non-null   float64
 31  Total Asset Return Growth Rate Ratio                    6819 non-null   float64
 32  Cash Reinvestment %                                     6819 non-null   float64
 33  Current Ratio                                           6819 non-null   float64
 34  Quick Ratio                                             6819 non-null   float64
 35  Interest Expense Ratio                                  6819 non-null   float64
```

```
36  Total debt/Total net worth                          6819 non-null   float64
37  Debt ratio %                                        6819 non-null   float64
38  Net worth/Assets                                    6819 non-null   float64
39  Long-term fund suitability ratio (A)                6819 non-null   float64
40  Borrowing dependency                                6819 non-null   float64
41  Contingent liabilities/Net worth                    6819 non-null   float64
42  Operating profit/Paid-in capital                    6819 non-null   float64
43  Net profit before tax/Paid-in capital               6819 non-null   float64
44  Inventory and accounts receivable/Net value         6819 non-null   float64
45  Total Asset Turnover                                6819 non-null   float64
46  Accounts Receivable Turnover                        6819 non-null   float64
47  Average Collection Days                             6819 non-null   float64
48  Inventory Turnover Rate (times)                     6819 non-null   float64
49  Fixed Assets Turnover Frequency                     6819 non-null   float64
50  Net Worth Turnover Rate (times)                     6819 non-null   float64
51  Revenue per person                                  6819 non-null   float64
52  Operating profit per person                         6819 non-null   float64
53  Allocation rate per person                          6819 non-null   float64
54  Working Capital to Total Assets                     6819 non-null   float64
55  Quick Assets/Total Assets                           6819 non-null   float64
56  Current Assets/Total Assets                         6819 non-null   float64
57  Cash/Total Assets                                   6819 non-null   float64
58  Quick Assets/Current Liability                      6819 non-null   float64
59  Cash/Current Liability                              6819 non-null   float64
60  Current Liability to Assets                         6819 non-null   float64
61  Operating Funds to Liability                        6819 non-null   float64
62  Inventory/Working Capital                           6819 non-null   float64
63  Inventory/Current Liability                         6819 non-null   float64
64  Current Liabilities/Liability                       6819 non-null   float64
65  Working Capital/Equity                              6819 non-null   float64
66  Current Liabilities/Equity                          6819 non-null   float64
67  Long-term Liability to Current Assets               6819 non-null   float64
68  Retained Earnings to Total Assets                   6819 non-null   float64
69  Total income/Total expense                          6819 non-null   float64
70  Total expense/Assets                                6819 non-null   float64
71  Current Asset Turnover Rate                         6819 non-null   float64
72  Quick Asset Turnover Rate                           6819 non-null   float64
73  Working capitcal Turnover Rate                      6819 non-null   float64
74  Cash Turnover Rate                                  6819 non-null   float64
75  Cash Flow to Sales                                  6819 non-null   float64
76  Fixed Assets to Assets                              6819 non-null   float64
77  Current Liability to Liability                      6819 non-null   float64
```

```
 78  Current Liability to Equity                              6819 non-null   float64
 79  Equity to Long-term Liability                            6819 non-null   float64
 80  Cash Flow to Total Assets                                6819 non-null   float64
 81  Cash Flow to Liability                                   6819 non-null   float64
 82  CFO to Assets                                            6819 non-null   float64
 83  Cash Flow to Equity                                      6819 non-null   float64
 84  Current Liability to Current Assets                      6819 non-null   float64
 85  Liability-Assets Flag                                    6819 non-null   int64
 86  Net Income to Total Assets                               6819 non-null   float64
 87  Total assets to GNP price                                6819 non-null   float64
 88  No-credit Interval                                       6819 non-null   float64
 89  Gross Profit to Sales                                    6819 non-null   float64
 90  Net Income to Stockholder's Equity                       6819 non-null   float64
 91  Liability to Equity                                      6819 non-null   float64
 92  Degree of Financial Leverage (DFL)                       6819 non-null   float64
 93  Interest Coverage Ratio (Interest expense to EBIT)       6819 non-null   float64
 94  Net Income Flag                                          6819 non-null   int64
 95  Equity to Liability                                      6819 non-null   float64
dtypes: float64(93), int64(3)
memory usage: 5.0 MB
```

In [9]:
```
1 df.dtypes
```

Out[9]:
```
Bankrupt?                                                   int64
ROA(C) before interest and depreciation before interest    float64
ROA(A) before interest and % after tax                     float64
ROA(B) before interest and depreciation after tax          float64
Operating Gross Margin                                     float64
                                                            ...
Liability to Equity                                        float64
Degree of Financial Leverage (DFL)                         float64
Interest Coverage Ratio (Interest expense to EBIT)         float64
Net Income Flag                                             int64
Equity to Liability                                        float64
Length: 96, dtype: object
```

In [10]:
```
1 df.describe()
```

Out[10]:

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industr income an expenditure/revenu
count	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.00000
mean	0.032263	0.505180	0.558625	0.553589	0.607948	0.607929	0.998755	0.797190	0.809084	0.30362
std	0.176710	0.060686	0.065620	0.061595	0.016934	0.016916	0.013010	0.012869	0.013601	0.01116
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000
25%	0.000000	0.476527	0.535543	0.527277	0.600445	0.600434	0.998969	0.797386	0.809312	0.30346
50%	0.000000	0.502706	0.559802	0.552278	0.605997	0.605976	0.999022	0.797464	0.809375	0.30352
75%	0.000000	0.535563	0.589157	0.584105	0.613914	0.613842	0.999095	0.797579	0.809469	0.30358
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.00000

8 rows × 96 columns

In [11]:
```
for col in df.describe(include='float').columns:
    print(col)
    print(df[col].unique ())
    print('*'*100)
```

```
ROA(C) before interest and depreciation before interest
[0.37059426 0.46429094 0.42607127 ... 0.47433335 0.54789646 0.49953688]
****************************************************************************************************
ROA(A) before interest and % after tax
[0.42438945 0.53821413 0.49901875 ... 0.49329481 0.6067379  0.6480048 ]
****************************************************************************************************
ROA(B) before interest and depreciation after tax
[0.40574977 0.51673002 0.47229509 ... 0.63172547 0.52417153 0.52063815]
****************************************************************************************************
Operating Gross Margin
[0.60145721 0.61023509 0.60145001 ... 0.60445524 0.61044408 0.62740887]
****************************************************************************************************
Realized Sales Gross Margin
[0.60145721 0.61023509 0.60136353 ... 0.60601191 0.61316825 0.62740887]
****************************************************************************************************
Operating Profit Rate
[0.9989692  0.99894598 0.99885735 ... 0.998741   0.99917649 0.9980803 ]
****************************************************************************************************
Pre-tax net Interest Rate
```

In [12]:
```
for col in df.describe(include='int').columns:
    print(col)
    print(df[col].unique ())
    print('*'*50)
```

```
Bankrupt?
[1 0]
**************************************************
Liability-Assets Flag
[0 1]
**************************************************
Net Income Flag
[1]
**************************************************
```

In [13]:
```
1 df.isnull().sum()
```

Out[13]:
```
Bankrupt?                                                 0
ROA(C) before interest and depreciation before interest   0
ROA(A) before interest and % after tax                    0
ROA(B) before interest and depreciation after tax         0
Operating Gross Margin                                    0
                                                         ..
Liability to Equity                                       0
Degree of Financial Leverage (DFL)                        0
Interest Coverage Ratio (Interest expense to EBIT)        0
Net Income Flag                                           0
Equity to Liability                                       0
Length: 96, dtype: int64
```

In [14]:

```python
def identify_and_impute_outliers(df, response_column):
    for column in df.columns:
        if column == response_column:
            continue
        if df[column].dtype in ['int64', 'float64']:  # Only consider numeric columns
            Q1 = df[column].quantile(0.25)
            Q3 = df[column].quantile(0.75)
            IQR = Q3 - Q1

            lower_bound = Q1 - 1.5 * IQR
            upper_bound = Q3 + 1.5 * IQR
            outliers = df[(df[column] < lower_bound) | (df[column] > upper_bound)]
            print(f"Column: {column}, Outliers Count: {len(outliers)}")
            median = df[column].median()
            df.loc[(df[column] < lower_bound) | (df[column] > upper_bound), column] = median

    return df
response_column = 'Bankrupt?'
df = identify_and_impute_outliers(df, response_column)
```

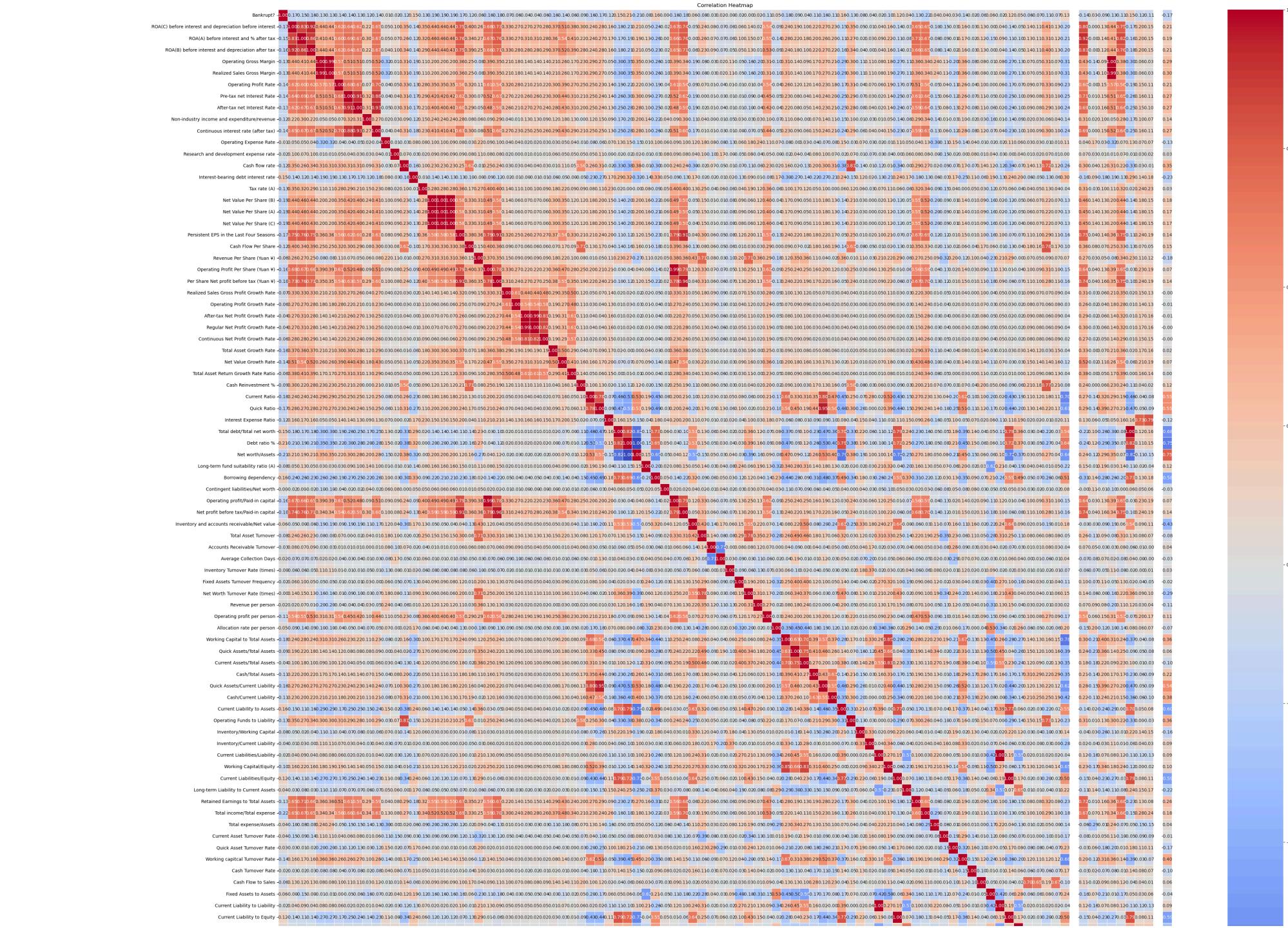
```
Column: ROA(C) before interest and depreciation before interest, Outliers Count: 391
Column: ROA(A) before interest and % after tax, Outliers Count: 561
Column: ROA(B) before interest and depreciation after tax, Outliers Count: 432
Column: Operating Gross Margin, Outliers Count: 320
Column: Realized Sales Gross Margin, Outliers Count: 318
Column: Operating Profit Rate, Outliers Count: 716
Column: Pre-tax net Interest Rate, Outliers Count: 773
Column: After-tax net Interest Rate, Outliers Count: 867
Column: Non-industry income and expenditure/revenue, Outliers Count: 1094
Column: Continuous interest rate (after tax), Outliers Count: 806
Column: Operating Expense Rate, Outliers Count: 0
Column: Research and development expense rate, Outliers Count: 182
Column: Cash flow rate, Outliers Count: 576
Column: Interest-bearing debt interest rate, Outliers Count: 396
Column: Tax rate (A), Outliers Count: 120
Column: Net Value Per Share (B), Outliers Count: 457
Column: Net Value Per Share (A), Outliers Count: 464
Column: Net Value Per Share (C), Outliers Count: 465
Column: Persistent EPS in the Last Four Seasons, Outliers Count: 508
Column: Cash Flow Per Share, Outliers Count: 532
Column: Revenue Per Share (Yuan ¥), Outliers Count: 478
Column: Operating Profit Per Share (Yuan ¥), Outliers Count: 442
Column: Per Share Net profit before tax (Yuan ¥), Outliers Count: 511
Column: Realized Sales Gross Profit Growth Rate, Outliers Count: 814
Column: Operating Profit Growth Rate, Outliers Count: 1008
Column: After-tax Net Profit Growth Rate, Outliers Count: 1033
Column: Regular Net Profit Growth Rate, Outliers Count: 1030
Column: Continuous Net Profit Growth Rate, Outliers Count: 1042
Column: Total Asset Growth Rate, Outliers Count: 1381
Column: Net Value Growth Rate, Outliers Count: 792
Column: Total Asset Return Growth Rate Ratio, Outliers Count: 674
Column: Cash Reinvestment %, Outliers Count: 617
Column: Current Ratio, Outliers Count: 589
Column: Quick Ratio, Outliers Count: 591
Column: Interest Expense Ratio, Outliers Count: 1362
Column: Total debt/Total net worth, Outliers Count: 407
Column: Debt ratio %, Outliers Count: 30
Column: Net worth/Assets, Outliers Count: 30
Column: Long-term fund suitability ratio (A), Outliers Count: 810
Column: Borrowing dependency, Outliers Count: 321
Column: Contingent liabilities/Net worth, Outliers Count: 942
```

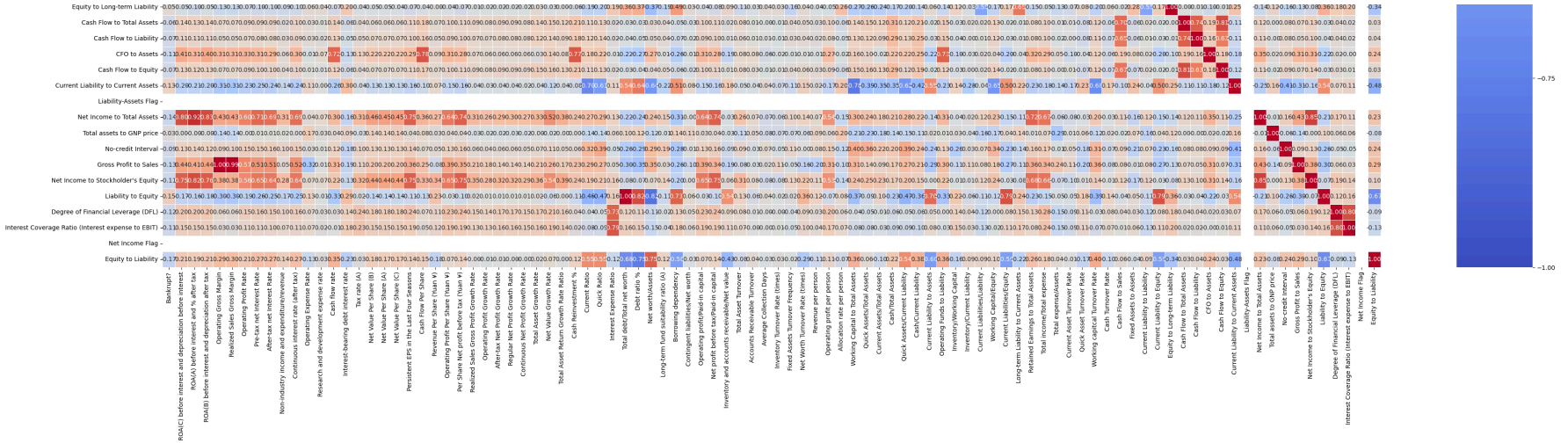
```
Column: Operating profit/Paid-in capital, Outliers Count: 446
Column: Net profit before tax/Paid-in capital, Outliers Count: 476
Column: Inventory and accounts receivable/Net value, Outliers Count: 421
Column: Total Asset Turnover, Outliers Count: 351
Column: Accounts Receivable Turnover, Outliers Count: 659
Column: Average Collection Days, Outliers Count: 193
Column: Inventory Turnover Rate (times), Outliers Count: 0
Column: Fixed Assets Turnover Frequency, Outliers Count: 1418
Column: Net Worth Turnover Rate (times), Outliers Count: 513
Column: Revenue per person, Outliers Count: 729
Column: Operating profit per person, Outliers Count: 876
Column: Allocation rate per person, Outliers Count: 693
Column: Working Capital to Total Assets, Outliers Count: 75
Column: Quick Assets/Total Assets, Outliers Count: 2
Column: Current Assets/Total Assets, Outliers Count: 0
Column: Cash/Total Assets, Outliers Count: 496
Column: Quick Assets/Current Liability, Outliers Count: 596
Column: Cash/Current Liability, Outliers Count: 728
Column: Current Liability to Assets, Outliers Count: 95
Column: Operating Funds to Liability, Outliers Count: 657
Column: Inventory/Working Capital, Outliers Count: 944
Column: Inventory/Current Liability, Outliers Count: 426
Column: Current Liabilities/Liability, Outliers Count: 40
Column: Working Capital/Equity, Outliers Count: 153
Column: Current Liabilities/Equity, Outliers Count: 480
Column: Long-term Liability to Current Assets, Outliers Count: 620
Column: Retained Earnings to Total Assets, Outliers Count: 633
Column: Total income/Total expense, Outliers Count: 463
Column: Total expense/Assets, Outliers Count: 372
Column: Current Asset Turnover Rate, Outliers Count: 1399
Column: Quick Asset Turnover Rate, Outliers Count: 0
Column: Working capitcal Turnover Rate, Outliers Count: 578
Column: Cash Turnover Rate, Outliers Count: 0
Column: Cash Flow to Sales, Outliers Count: 1052
Column: Fixed Assets to Assets, Outliers Count: 62
Column: Current Liability to Liability, Outliers Count: 40
Column: Current Liability to Equity, Outliers Count: 480
Column: Equity to Long-term Liability, Outliers Count: 406
Column: Cash Flow to Total Assets, Outliers Count: 878
Column: Cash Flow to Liability, Outliers Count: 1212
Column: CFO to Assets, Outliers Count: 342
Column: Cash Flow to Equity, Outliers Count: 827
```

```
Column: Current Liability to Current Assets, Outliers Count: 276
Column: Liability-Assets Flag, Outliers Count: 8
Column: Net Income to Total Assets, Outliers Count: 561
Column: Total assets to GNP price, Outliers Count: 797
Column: No-credit Interval, Outliers Count: 1139
Column: Gross Profit to Sales, Outliers Count: 320
Column: Net Income to Stockholder's Equity, Outliers Count: 571
Column: Liability to Equity, Outliers Count: 404
Column: Degree of Financial Leverage (DFL), Outliers Count: 1503
Column: Interest Coverage Ratio (Interest expense to EBIT), Outliers Count: 1421
Column: Net Income Flag, Outliers Count: 0
Column: Equity to Liability, Outliers Count: 549
```

In [17]:

```
numeric_columns = df.select_dtypes(include=['int64', 'float64'])

correlation_matrix = numeric_columns.corr()
plt.figure(figsize=(45, 45))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f",linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```


Correlation Heatmap
Bankrupt?
ROA(C) before interest and depreciation before interest
ROA(A) before interest and % after tax
ROA(B) before interest and depreciation after tax
Operating Gross Margin
Realized Sales Gross Margin
Operating Profit Rate
Pre-tax net Interest Rate
After-tax net Interest Rate
Non-industry income and expenditure/revenue
Continuous interest rate (after tax)
Operating Expense Rate
Research and development expense rate
Cash flow rate
Interest-bearing debt interest rate
Tax rate (A)
Net Value Per Share (B)
Net Value Per Share (A)
Net Value Per Share (C)
Persistent EPS in the Last Four Seasons
Cash Flow Per Share
Revenue Per Share (Yuan ¥)
Operating Profit Per Share (Yuan ¥)
Per Share Net profit before tax (Yuan ¥)
Realized Sales Gross Profit Growth Rate
Operating Profit Growth Rate
After-tax Net Profit Growth Rate
Regular Net Profit Growth Rate
Continuous Net Profit Growth Rate
Total Asset Growth Rate
Net Value Growth Rate
Total Asset Return Growth Rate Ratio
Cash Reinvestment %
Current Ratio
Quick Ratio
Interest Expense Ratio
Total debt/Total net worth
Debt ratio %
Net worth/Assets
Long-term fund suitability ratio (A)
Borrowing dependency
Contingent liabilities/Net worth
Operating profit/Paid-in capital
Net profit before tax/Paid-in capital
Inventory and accounts receivable/Net value
Total Asset Turnover
Accounts Receivable Turnover
Average Collection Days
Inventory Turnover Rate (times)
Fixed Assets Turnover Frequency
Net Worth Turnover Rate (times)
Revenue per person
Operating profit per person
Allocation rate per person
Working Capital to Total Assets
Quick Assets/Total Assets
Current Assets/Total Assets
Cash/Total Assets
Quick Assets/Current Liability
Cash/Current Liability
Current Liability to Assets
Operating Funds to Liability
Inventory/Working Capital
Inventory/Current Liability
Current Liabilities/Liability
Working Capital/Equity
Current Liabilities/Equity
Long-term Liability to Current Assets
Retained Earnings to Total Assets
Total income/Total expense
Total expense/Assets
Current Asset Turnover Rate
Quick Asset Turnover Rate
Working capital Turnover Rate
Cash Turnover Rate
Cash Flow to Sales
Fixed Assets to Assets
Current Liability to Liability
Current Liability to Equity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50



In [18]:
```
correlation_matrix = df.corr()
get = (correlation_matrix['Bankrupt?']>0.20) | (correlation_matrix['Bankrupt?']<-0.20)
get.sum()
pred_colms = (get[get].index.to_list())
pred_colms

new_df = df[pred_colms]
new_df.shape
```

Out[18]: (6819, 4)

In [20]:
```
1 new_df
```

Out[20]:

	Bankrupt?	Debt ratio %	Net worth/Assets	Total income/Total expense
0	1	0.207576	0.792424	0.002022
1	1	0.171176	0.828824	0.002226
2	1	0.207516	0.792484	0.002060
3	1	0.151465	0.848535	0.002336
4	1	0.106509	0.893491	0.002224
...	...	...	...	...
6814	0	0.124618	0.875382	0.002266
6815	0	0.099253	0.900747	0.002288
6816	0	0.038939	0.961061	0.002239
6817	0	0.086979	0.913021	0.002395
6818	0	0.014149	0.985851	0.002791

6819 rows × 4 columns

In [24]:
```
1 numeric_categorical_df = df[['Bankrupt?','Liability-Assets Flag','Net Income Flag']]
```

In [25]:
```
1 numeric_categorical_df
```

Out[25]:

	Bankrupt?	Liability-Assets Flag	Net Income Flag
0	1	0	1
1	1	0	1
2	1	0	1
3	1	0	1
4	1	0	1
...	...	...	...
6814	0	0	1
6815	0	0	1
6816	0	0	1
6817	0	0	1
6818	0	0	1

6819 rows × 3 columns

In [28]:
```
1 #H0bl There is no association between the 'Bankrupt?' status of a company and its 'Liability-Assets Flag'
2 contingency_table_liability = pd.crosstab(df['Bankrupt?'], df['Liability-Assets Flag'])
3 chi2_stat_la, p_value_la, dof_la, expected_freq_la = chi2_contingency(contingency_table_liability)
4 #H0bn  There is no association between the 'Bankrupt?' status of a company and its 'Net Income Flag'.
5 contingency_table_ni = pd.crosstab(df['Bankrupt?'], df['Net Income Flag'])
6 chi2_stat_ni, p_value_ni, dof_ni, expected_freq_ni = chi2_contingency(contingency_table_ni)
```

In [30]:
```
1 print(str(p_value_la)+" "+str(p_value_ni))
```

```
1.0 1.0
```

In [29]:
```
# based on above chi-square test, it has been found that numeric_categorical data has no effect on the the respon
# so for analysis part i am not using those 2 columns

# Based on corretion matrix method we have found only 3 columns which are effective immpact on response variable
#although more threshold value less data set columns will lead to strong prediction model i will use only 0.2 as

# In one way i am doing feature selection for numeric and catgorical data which is not as per proceess
# But instead of doing EDA/ visualization on large amount of predictors, it will be good to perform visualization

#But i am sure that remaining columns also affect respose variable.
# So i am conclude that i am doing this analysis based on very little columns and it may be strong prediction
#in coming future i suggest do analysis based on domain knowledge
```

In [31]:
```
new_df.columns=['Bankrupt','Debt_ratio','Net_worth_Assets','income_Expense_Ratio']
```

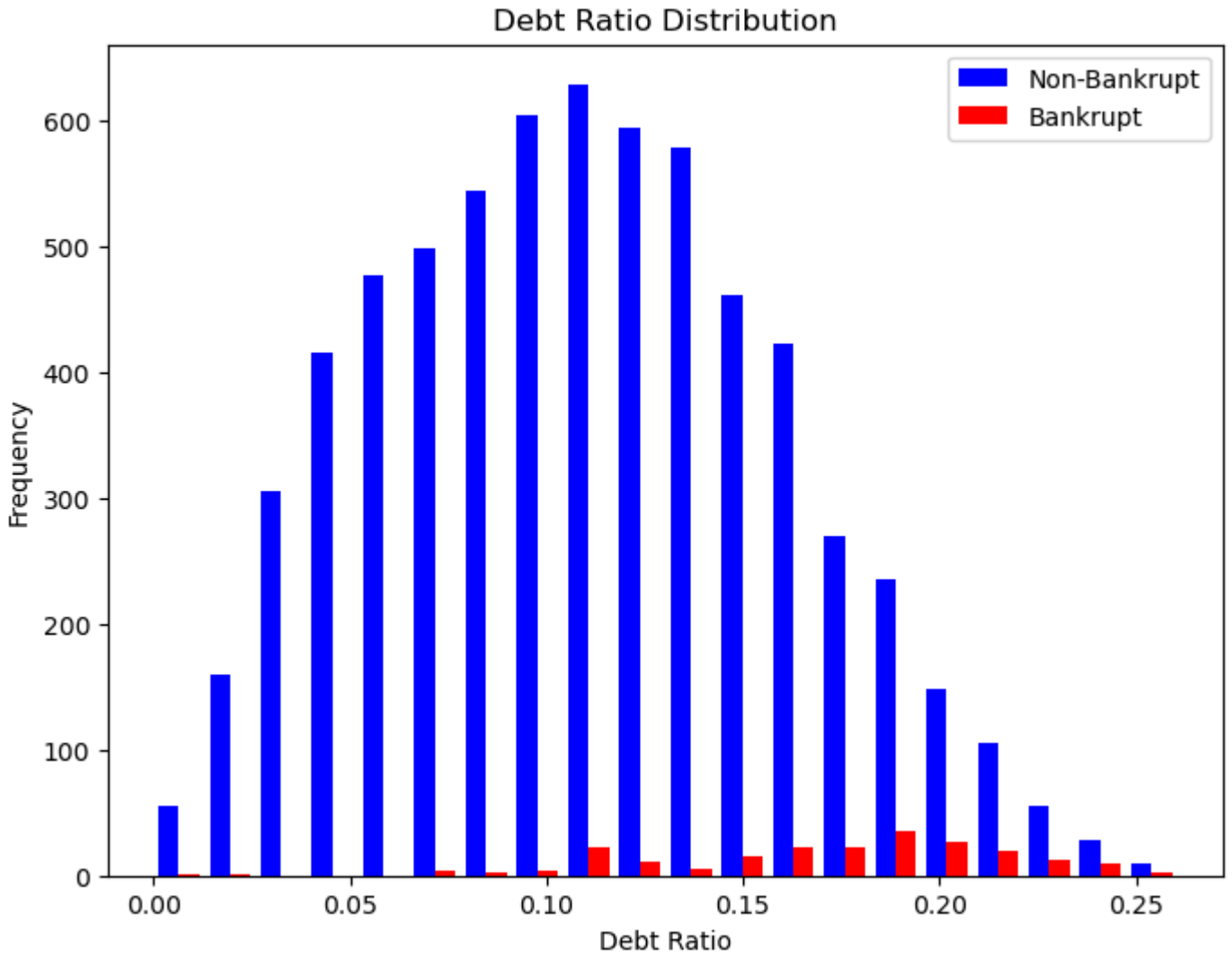
In [45]:
```
new_df.head()
```

Out[45]:

	Bankrupt	Debt_ratio	Net_worth_Assets	income_Expense_Ratio
0	1	0.207576	0.792424	0.002022
1	1	0.171176	0.828824	0.002226
2	1	0.207516	0.792484	0.002060
3	1	0.151465	0.848535	0.002336
4	1	0.106509	0.893491	0.002224

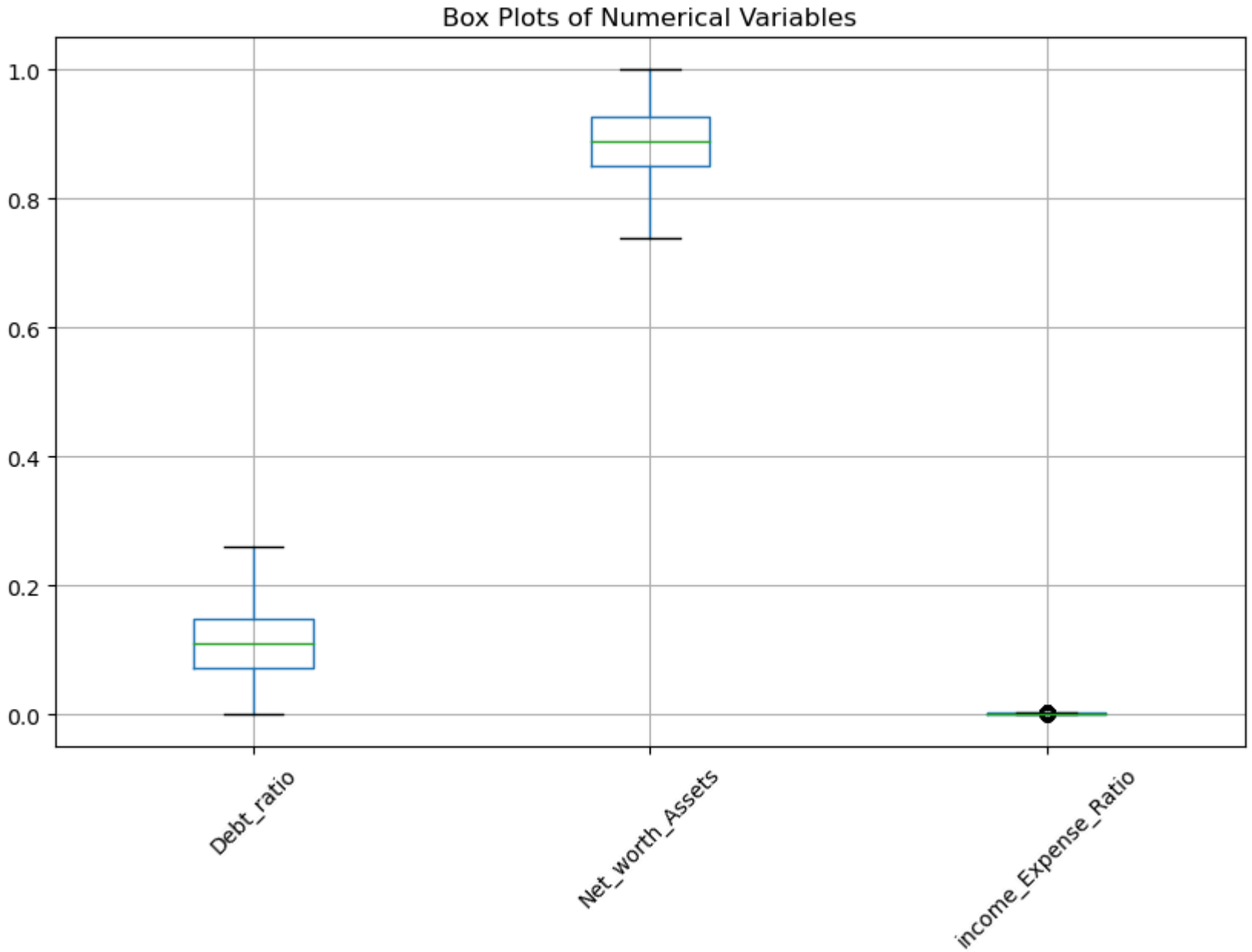
In [39]:

```
plt.figure(figsize=(8, 6))
plt.hist([new_df[new_df['Bankrupt'] == 0]['Debt_ratio'], new_df[new_df['Bankrupt'] == 1]['Debt_ratio']], bins=20,
plt.title('Debt Ratio Distribution')
plt.xlabel('Debt Ratio')
plt.ylabel('Frequency')
plt.legend()
plt.show()

```

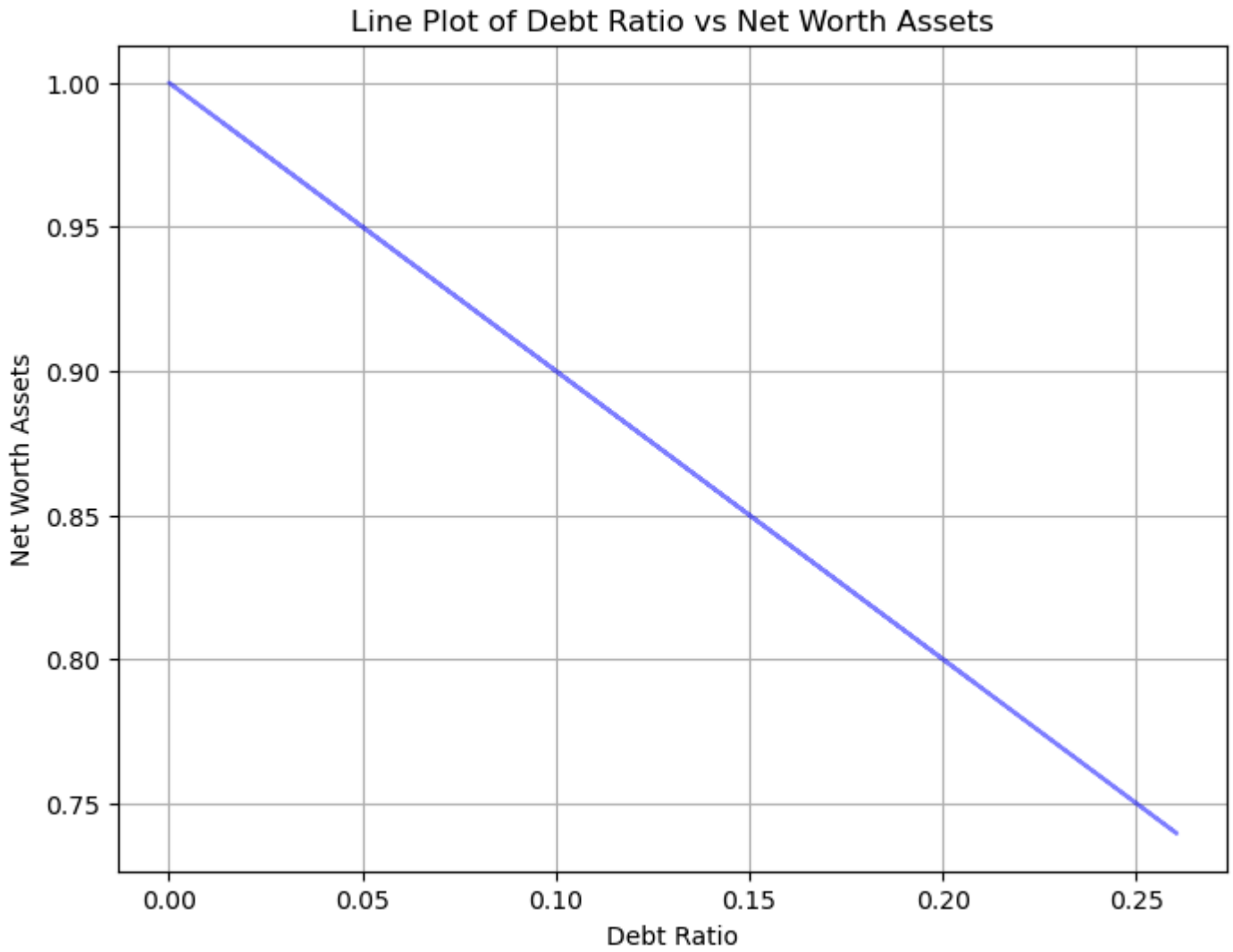

Debt Ratio Distribution
Non-Bankrupt
Bankrupt
Frequency
0
100
200
300
400
500
600
0.00
0.05
0.10
0.15
0.20
0.25
Debt Ratio

In [35]:

```
plt.figure(figsize=(10, 6))
new_df.iloc[:, 1:].boxplot()
plt.title('Box Plots of Numerical Variables')
plt.xticks(rotation=45)
plt.show()

```

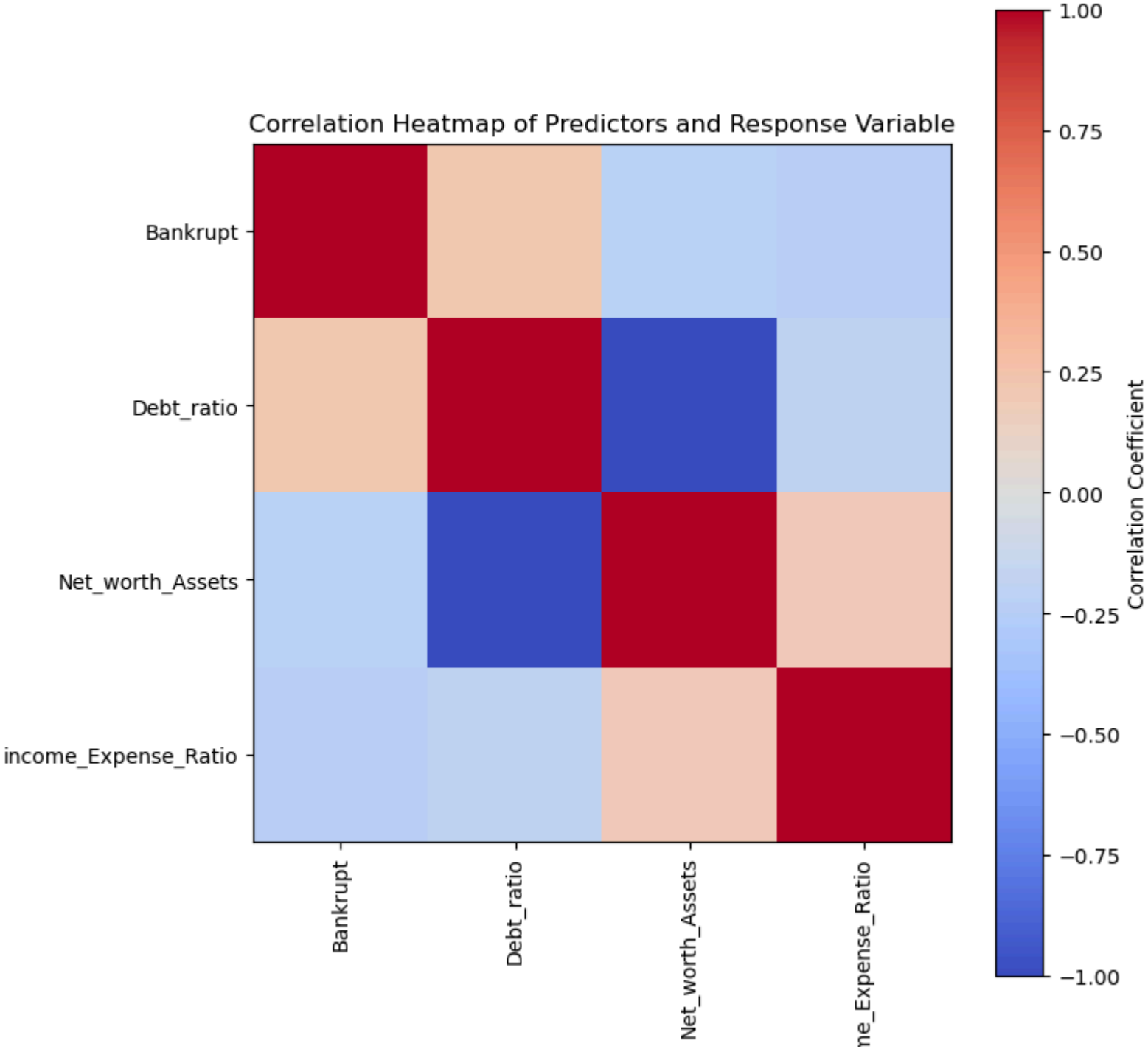

Box Plots of Numerical Variables
1.0
0.8
0.6
0.4
0.2
0.0
Debt_ratio
Net_worth_Assets
income_Expense_Ratio

In [41]:
```
plt.figure(figsize=(8, 6))
plt.plot(new_df['Debt_ratio'], new_df['Net_worth_Assets'], color='blue', alpha=0.5)
plt.title('Line Plot of Debt Ratio vs Net Worth Assets')
plt.xlabel('Debt Ratio')
plt.ylabel('Net Worth Assets')
plt.grid(True)
plt.show()
```



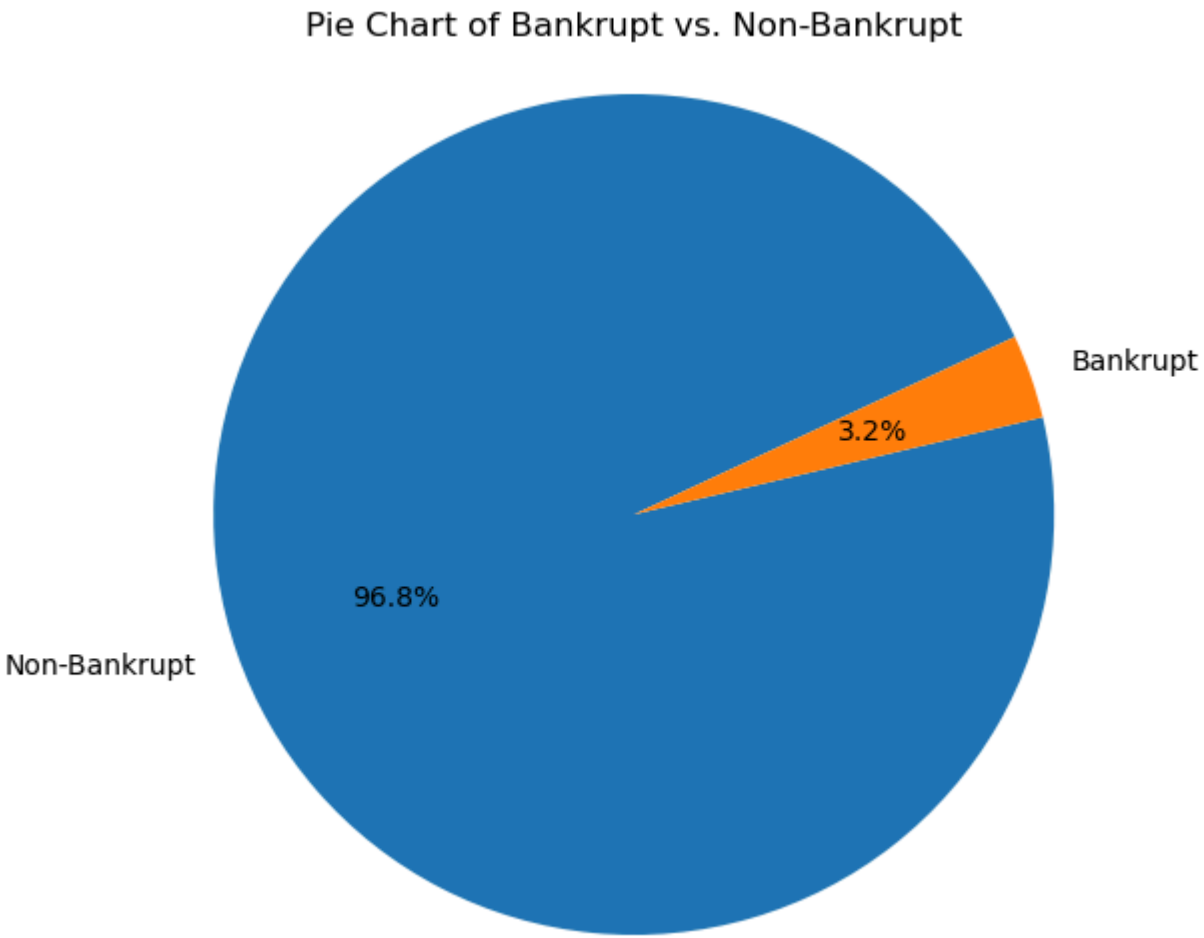
In [37]:

```
corr_matrix = new_df.corr()
plt.figure(figsize=(8, 8))
plt.imshow(corr_matrix, cmap='coolwarm', interpolation='nearest')
plt.colorbar(label='Correlation Coefficient')
plt.title('Correlation Heatmap of Predictors and Response Variable')
plt.xticks(range(len(corr_matrix)), corr_matrix.columns, rotation=90)
plt.yticks(range(len(corr_matrix)), corr_matrix.columns)
plt.tight_layout()
plt.show()
```


Correlation Heatmap of Predictors and Response Variable
Bankrupt
Debt_ratio
Net_worth_Assets
income_Expense_Ratio
Bankrupt
Debt_ratio
Net_worth_Assets
ne_Expense_Ratio
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Correlation Coefficient

incor

In [56]:
```
bankrupt_counts = new_df['Bankrupt'].value_counts()
plt.figure(figsize=(8, 6))
plt.pie(bankrupt_counts, labels=['Non-Bankrupt', 'Bankrupt'], autopct='%1.1f%%',startangle=25)
plt.title('Pie Chart of Bankrupt vs. Non-Bankrupt')
plt.axis('equal')
plt.show()

```



In [83]:
```
x = new_df[['Debt_ratio', 'Net_worth_Assets', 'income_Expense_Ratio']]
y = new_df['Bankrupt']
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
x_train=sm.add_constant(x_train,prepend=False)
x_test=sm.add_constant(x_test,prepend=False)
model = sm.MNLogit(y_train, x_train).fit()
y_pred = model.predict(x_test)
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.096414
         Iterations: 35
```

In [84]:
```
print(model.summary())
```

```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:                Bankrupt   No. Observations:                 5455
Model:                         MNLogit   Df Residuals:                     5451
Method:                            MLE   Df Model:                            3
Date:                 Mon, 17 Jun 2024   Pseudo R-squ.:                  0.3020
Time:                         21:06:57   Log-Likelihood:                -525.94
converged:                       False   LL-Null:                       -753.53
Covariance Type:             nonrobust   LLR p-value:                 2.463e-98
=============================================================================================
          Bankrupt=1       coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Debt_ratio              17.2469   5.21e+06   3.31e-06      1.000   -1.02e+07    1.02e+07
Net_worth_Assets        -3.3807   5.21e+06  -6.49e-07      1.000   -1.02e+07    1.02e+07
income_Expense_Ratio -7584.9310    577.352    -13.137      0.000   -8716.520   -6453.342
const                   13.8584   5.21e+06   2.66e-06      1.000   -1.02e+07    1.02e+07
=============================================================================================
```

In [ ]:
```
1. Debt Ratio: The coefficient is very high (17.9178), but the p-value is very high (1.000),
indicating that the variable is not statistically significant in predicting bankruptcy. Therefore,
we fail to reject the null hypothesis that the debt ratio has no effect on bankruptcy.

2. Net Worth Assets: The coefficient is negative (-3.4177), but like the debt ratio, the p-value is very high (1.
indicating that this variable is also not statistically significant in predicting bankruptcy.

3. Income Expense Ratio: The coefficient is quite large in magnitude (-7924.3796),
and the p-value is very low (close to 0), indicating that this variable is statistically significant in predictir
Therefore, we reject the null hypothesis that the income expense ratio has no effect on bankruptcy.

Based on this analysis, the income expense ratio seems to be the most influential predictor of bankruptcy among t

A higher income expense ratio means a company is spending more compared to its income, which could indicate finar
increase the likelihood of bankruptcy.
This is evident from its substantial coefficient value (-7924.3796), indicating a strong impact on the likelihood
Therefore, understanding and managing this ratio is important for avoiding financial distress.
```

In [ ]:
```
Debt Ratio (Positive Coefficient): With a coefficient of approximately 17.92, an increase in the debt ratio sugge
  bankruptcy. For each unit increase in the debt ratio, the probability of bankruptcy rises.

2. Net Worth Assets (Negative Coefficient): The coefficient of around -3.42 implies that a decrease in net worth
  bankruptcy. Each unit decrease in net worth assets corresponds to a higher probability of facing bankruptcy.

3. Income Expense Ratio (Negative Coefficient): With a coefficient of about -7924.38, an increase in the income
   expense ratio indicates a reduced risk of bankruptcy. For every unit increase in this ratio, the probability c
```

In [86]:
```
y_pred = model.predict(x_test)
```

In [87]:
```
1 y_pred.head()
```

Out[87]:

	0	1
239	0.985231	0.014769
2850	0.993688	0.006312
2687	0.998043	0.001957
6500	0.748266	0.251734
2684	0.984981	0.015019

In [88]:
```
1 y_test.head()
```

Out[88]:
```
239     0
2850    0
2687    0
6500    1
2684    0
Name: Bankrupt, dtype: int64
```

In [75]:
```
1 y_pred_class = y_pred.idxmax(axis=1)
```

In [76]:
```
1 precision = precision_score(y_test, y_pred_class)
2 recall = recall_score(y_test, y_pred_class)
3 accuracy = accuracy_score(y_test, y_pred_class)
4 f1 = f1_score(y_test, y_pred_class)
5 conf_matrix = confusion_matrix(y_test, y_pred_class)
```

In [77]:
```
1 accuracy
```

Out[77]:
```
0.9618768328445748
```

In [78]:
```
1 conf_matrix
```

Out[78]:
```
array([[1309,    4],
       [  48,    3]], dtype=int64)
```

In [79]:
```
1 precision
```

Out[79]:
```
0.42857142857142855
```

In [80]:
```
1 recall
```

Out[80]:
```
0.058823529411764705
```

In [81]:
```
1 f1
```

Out[81]:
```
0.10344827586206895
```

In [ ]:
```
1  The precision score of 0.4285 indicates that when the model predicts bankruptcy, it is correct approximately 42.8
2  The recall score of 0.058 indicates that the model correctly identifies approximately 5.8% of the actual bankrupt
3
4  The high accuracy score of 0.960 suggests that the model performs well overall, but the low precision and recall
5  indicate that it may struggle to correctly identify bankruptcies, which could be due to class imbalance or other
6
7  Based on the confusion matrix:
8
9  True Negatives (TN): 1309
10 False Positives (FP): 4
11 False Negatives (FN): 48
12 True Positives (TP): 3
13 This matrix illustrates the model's performance in classifying instances as bankrupt or non-bankrupt.
14 It shows that the model correctly identified 1309 instances as non-bankrupt (TN) and 3 instances as bankrupt (TP)
15 However, it incorrectly classified 6 instances as bankrupt when they were actually non-bankrupt (FP), and 49 inst
16 were actually bankrupt (FN).
```

In [ ]:
```
so in summary

The model achieved a high accuracy of 96.18%, indicating that it correctly classified the majority of instances.

However, the precision (42.85%) and recall (5.88%) for bankrupt instances are low, suggesting that the model's ab
to correctly identify bankrupt cases is limited.

The confusion matrix reveals that while the model correctly identified a large number of non-bankrupt instances,
it struggled with correctly classifying bankrupt instances, resulting in a higher number of
false negatives and false positives.
```

In [ ]:
```

```

In [ ]:
```

```