Гусев Сергей, ИУ5Ц-82Б метод №1 - Метод опорных векторов; метод №2 - Случайный лес.

0. Подготовка

In [1]:

```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer
from sklearn.impute import MissingIndicator
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import balanced_accuracy_score, plot_roc_curve, confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
```

In [2]:

```
# отбираем 5000 строк из всего датасета
data = pd.read_csv('data.csv', nrows=5000)
```

In [3]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 96 columns):
 #   Column                                                  Non-Null Count  Dtype  
---  ------                                                  --------------  -----  
 0   Bankrupt?                                               5000 non-null   int64  
 1    ROA(C) before interest and depreciation before interest 5000 non-null   float64
 2    ROA(A) before interest and % after tax                  5000 non-null   float64
 3    ROA(B) before interest and depreciation after tax       5000 non-null   float64
 4    Operating Gross Margin                                  5000 non-null   float64
 5    Realized Sales Gross Margin                             5000 non-null   float64
 6    Operating Profit Rate                                   5000 non-null   float64
 7    Pre-tax net Interest Rate                               5000 non-null   float64
 8    After-tax net Interest Rate                             5000 non-null   float64
 9    Non-industry income and expenditure/revenue             5000 non-null   float64
 10   Continuous interest rate (after tax)                    5000 non-null   float64
 11   Operating Expense Rate                                  5000 non-null   float64
 12   Research and development expense rate                   5000 non-null   float64
 13   Cash flow rate                                          5000 non-null   float64
 14   Interest-bearing debt interest rate                     5000 non-null   float64
 15   Tax rate (A)                                            5000 non-null   float64
 16   Net Value Per Share (B)                                 5000 non-null   float64
 17   Net Value Per Share (A)                                 5000 non-null   float64
 18   Net Value Per Share (C)                                 5000 non-null   float64
 19   Persistent EPS in the Last Four Seasons                 5000 non-null   float64
 20   Cash Flow Per Share                                     5000 non-null   float64
 21   Revenue Per Share (Yuan ¥)                              5000 non-null   float64
 22   Operating Profit Per Share (Yuan ¥)                     5000 non-null   float64
 23   Per Share Net profit before tax (Yuan ¥)                5000 non-null   float64
 24   Realized Sales Gross Profit Growth Rate                 5000 non-null   float64
 25   Operating Profit Growth Rate                            5000 non-null   float64
 26   After-tax Net Profit Growth Rate                        5000 non-null   float64
 27   Regular Net Profit Growth Rate                          5000 non-null   float64
 28   Continuous Net Profit Growth Rate                       5000 non-null   float64
 29   Total Asset Growth Rate                                 5000 non-null   float64
 30   Net Value Growth Rate                                   5000 non-null   float64
 31   Total Asset Return Growth Rate Ratio                    5000 non-null   float64
 32   Cash Reinvestment %                                     5000 non-null   float64
 33   Current Ratio                                           5000 non-null   float64
 34   Quick Ratio                                             5000 non-null   float64
 35   Interest Expense Ratio                                  5000 non-null   float64
 36   Total debt/Total net worth                              5000 non-null   float64
 37   Debt ratio %                                            5000 non-null   float64
 38   Net worth/Assets                                        5000 non-null   float64
 39   Long-term fund suitability ratio (A)                    5000 non-null   float64
 40   Borrowing dependency                                    5000 non-null   float64
 41   Contingent liabilities/Net worth                        5000 non-null   float64
 42   Operating profit/Paid-in capital                        5000 non-null   float64
 43   Net profit before tax/Paid-in capital                   5000 non-null   float64
 44   Inventory and accounts receivable/Net value             5000 non-null   float64
 45   Total Asset Turnover                                    5000 non-null   float64
 46   Accounts Receivable Turnover                            5000 non-null   float64
 47   Average Collection Days                                 5000 non-null   float64
 48   Inventory Turnover Rate (times)                         5000 non-null   float64
 49   Fixed Assets Turnover Frequency                         5000 non-null   float64
 50   Net Worth Turnover Rate (times)                         5000 non-null   float64
 51   Revenue per person                                      5000 non-null   float64
 52   Operating profit per person                             5000 non-null   float64
 53   Allocation rate per person                              5000 non-null   float64
 54   Working Capital to Total Assets                         5000 non-null   float64
 55   Quick Assets/Total Assets                               5000 non-null   float64
 56   Current Assets/Total Assets                             5000 non-null   float64
 57   Cash/Total Assets                                       5000 non-null   float64
 58   Quick Assets/Current Liability                          5000 non-null   float64
 59   Cash/Current Liability                                  5000 non-null   float64
 60   Current Liability to Assets                             5000 non-null   float64
 61   Operating Funds to Liability                            5000 non-null   float64
 62   Inventory/Working Capital                               5000 non-null   float64
 63   Inventory/Current Liability                             5000 non-null   float64
 64   Current Liabilities/Liability                           5000 non-null   float64
 65   Working Capital/Equity                                  5000 non-null   float64
```

```
 66   Current Liabilities/Equity                              5000 non-null   float64
 67   Long-term Liability to Current Assets                   5000 non-null   float64
 68   Retained Earnings to Total Assets                       5000 non-null   float64
 69   Total income/Total expense                              5000 non-null   float64
 70   Total expense/Assets                                    5000 non-null   float64
 71   Current Asset Turnover Rate                             5000 non-null   float64
 72   Quick Asset Turnover Rate                               5000 non-null   float64
 73   Working capitcal Turnover Rate                          5000 non-null   float64
 74   Cash Turnover Rate                                      5000 non-null   float64
 75   Cash Flow to Sales                                      5000 non-null   float64
 76   Fixed Assets to Assets                                  5000 non-null   float64
 77   Current Liability to Liability                          5000 non-null   float64
 78   Current Liability to Equity                             5000 non-null   float64
 79   Equity to Long-term Liability                           5000 non-null   float64
 80   Cash Flow to Total Assets                               5000 non-null   float64
 81   Cash Flow to Liability                                  5000 non-null   float64
 82   CFO to Assets                                           5000 non-null   float64
 83   Cash Flow to Equity                                     5000 non-null   float64
 84   Current Liability to Current Assets                     5000 non-null   float64
 85   Liability-Assets Flag                                   5000 non-null   int64
 86   Net Income to Total Assets                              5000 non-null   float64
 87   Total assets to GNP price                               5000 non-null   float64
 88   No-credit Interval                                      5000 non-null   float64
 89   Gross Profit to Sales                                   5000 non-null   float64
 90   Net Income to Stockholder's Equity                      5000 non-null   float64
 91   Liability to Equity                                     5000 non-null   float64
 92   Degree of Financial Leverage (DFL)                      5000 non-null   float64
 93   Interest Coverage Ratio (Interest expense to EBIT)      5000 non-null   float64
 94   Net Income Flag                                         5000 non-null   int64
 95   Equity to Liability                                     5000 non-null   float64
dtypes: float64(93), int64(3)
memory usage: 3.7 MB
```

In [5]:

```
# Оцениваем баланс классов целевого признака
data['Bankrupt?'].value_counts()/data['Bankrupt?'].shape[0]*100
```

Out[5]:

```
0    96.16
1     3.84
Name: Bankrupt?, dtype: float64
```

In [6]:

```
# Проверяем процент пропусков в данных для всех колонок
(data.isnull().sum()/data.shape[0]*100).sort_values(ascending=False)
```

Out[6]:

```
 Equity to Liability                    0.0
 Net Income Flag                        0.0
 Operating Profit Growth Rate           0.0
 After-tax Net Profit Growth Rate       0.0
 Regular Net Profit Growth Rate         0.0
                                       ...
 Current Liabilities/Equity             0.0
 Long-term Liability to Current Assets  0.0
 Retained Earnings to Total Assets      0.0
 Total income/Total expense             0.0
Bankrupt?                               0.0
Length: 96, dtype: float64
```

Пропусков нет, так что двигаемся дальше

In [8]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 96 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Bankrupt?                                               5000 non-null   int64
 1    ROA(C) before interest and depreciation before interest 5000 non-null   float64
 2    ROA(A) before interest and % after tax                  5000 non-null   float64
 3    ROA(B) before interest and depreciation after tax       5000 non-null   float64
 4    Operating Gross Margin                                  5000 non-null   float64
 5    Realized Sales Gross Margin                             5000 non-null   float64
 6    Operating Profit Rate                                   5000 non-null   float64
 7    Pre-tax net Interest Rate                               5000 non-null   float64
 8    After-tax net Interest Rate                             5000 non-null   float64
 9    Non-industry income and expenditure/revenue             5000 non-null   float64
 10   Continuous interest rate (after tax)                    5000 non-null   float64
 11   Operating Expense Rate                                  5000 non-null   float64
 12   Research and development expense rate                   5000 non-null   float64
 13   Cash flow rate                                          5000 non-null   float64
 14   Interest-bearing debt interest rate                     5000 non-null   float64
 15   Tax rate (A)                                            5000 non-null   float64
 16   Net Value Per Share (B)                                 5000 non-null   float64
 17   Net Value Per Share (A)                                 5000 non-null   float64
 18   Net Value Per Share (C)                                 5000 non-null   float64
 19   Persistent EPS in the Last Four Seasons                 5000 non-null   float64
 20   Cash Flow Per Share                                     5000 non-null   float64
 21   Revenue Per Share (Yuan ¥)                              5000 non-null   float64
 22   Operating Profit Per Share (Yuan ¥)                     5000 non-null   float64
 23   Per Share Net profit before tax (Yuan ¥)                5000 non-null   float64
 24   Realized Sales Gross Profit Growth Rate                 5000 non-null   float64
 25   Operating Profit Growth Rate                            5000 non-null   float64
 26   After-tax Net Profit Growth Rate                        5000 non-null   float64
 27   Regular Net Profit Growth Rate                          5000 non-null   float64
 28   Continuous Net Profit Growth Rate                       5000 non-null   float64
 29   Total Asset Growth Rate                                 5000 non-null   float64
 30   Net Value Growth Rate                                   5000 non-null   float64
 31   Total Asset Return Growth Rate Ratio                    5000 non-null   float64
 32   Cash Reinvestment %                                     5000 non-null   float64
```

```
 33   Current Ratio                                      5000 non-null   float64
 34   Quick Ratio                                        5000 non-null   float64
 35   Interest Expense Ratio                             5000 non-null   float64
 36   Total debt/Total net worth                         5000 non-null   float64
 37   Debt ratio %                                       5000 non-null   float64
 38   Net worth/Assets                                   5000 non-null   float64
 39   Long-term fund suitability ratio (A)               5000 non-null   float64
 40   Borrowing dependency                               5000 non-null   float64
 41   Contingent liabilities/Net worth                   5000 non-null   float64
 42   Operating profit/Paid-in capital                   5000 non-null   float64
 43   Net profit before tax/Paid-in capital              5000 non-null   float64
 44   Inventory and accounts receivable/Net value        5000 non-null   float64
 45   Total Asset Turnover                               5000 non-null   float64
 46   Accounts Receivable Turnover                       5000 non-null   float64
 47   Average Collection Days                            5000 non-null   float64
 48   Inventory Turnover Rate (times)                    5000 non-null   float64
 49   Fixed Assets Turnover Frequency                    5000 non-null   float64
 50   Net Worth Turnover Rate (times)                    5000 non-null   float64
 51   Revenue per person                                 5000 non-null   float64
 52   Operating profit per person                        5000 non-null   float64
 53   Allocation rate per person                         5000 non-null   float64
 54   Working Capital to Total Assets                    5000 non-null   float64
 55   Quick Assets/Total Assets                          5000 non-null   float64
 56   Current Assets/Total Assets                        5000 non-null   float64
 57   Cash/Total Assets                                  5000 non-null   float64
 58   Quick Assets/Current Liability                     5000 non-null   float64
 59   Cash/Current Liability                             5000 non-null   float64
 60   Current Liability to Assets                        5000 non-null   float64
 61   Operating Funds to Liability                       5000 non-null   float64
 62   Inventory/Working Capital                          5000 non-null   float64
 63   Inventory/Current Liability                        5000 non-null   float64
 64   Current Liabilities/Liability                      5000 non-null   float64
 65   Working Capital/Equity                             5000 non-null   float64
 66   Current Liabilities/Equity                         5000 non-null   float64
 67   Long-term Liability to Current Assets              5000 non-null   float64
 68   Retained Earnings to Total Assets                  5000 non-null   float64
 69   Total income/Total expense                         5000 non-null   float64
 70   Total expense/Assets                               5000 non-null   float64
 71   Current Asset Turnover Rate                        5000 non-null   float64
 72   Quick Asset Turnover Rate                          5000 non-null   float64
 73   Working capitcal Turnover Rate                     5000 non-null   float64
 74   Cash Turnover Rate                                 5000 non-null   float64
 75   Cash Flow to Sales                                 5000 non-null   float64
 76   Fixed Assets to Assets                             5000 non-null   float64
 77   Current Liability to Liability                     5000 non-null   float64
 78   Current Liability to Equity                        5000 non-null   float64
 79   Equity to Long-term Liability                      5000 non-null   float64
 80   Cash Flow to Total Assets                          5000 non-null   float64
 81   Cash Flow to Liability                             5000 non-null   float64
 82   CFO to Assets                                      5000 non-null   float64
 83   Cash Flow to Equity                                5000 non-null   float64
 84   Current Liability to Current Assets                5000 non-null   float64
 85   Liability-Assets Flag                              5000 non-null   int64
 86   Net Income to Total Assets                         5000 non-null   float64
 87   Total assets to GNP price                          5000 non-null   float64
 88   No-credit Interval                                 5000 non-null   float64
 89   Gross Profit to Sales                              5000 non-null   float64
 90   Net Income to Stockholder's Equity                 5000 non-null   float64
 91   Liability to Equity                                5000 non-null   float64
 92   Degree of Financial Leverage (DFL)                 5000 non-null   float64
 93   Interest Coverage Ratio (Interest expense to EBIT) 5000 non-null   float64
 94   Net Income Flag                                    5000 non-null   int64
 95   Equity to Liability                                5000 non-null   float64
dtypes: float64(93), int64(3)
memory usage: 3.7 MB
```

In [9]:

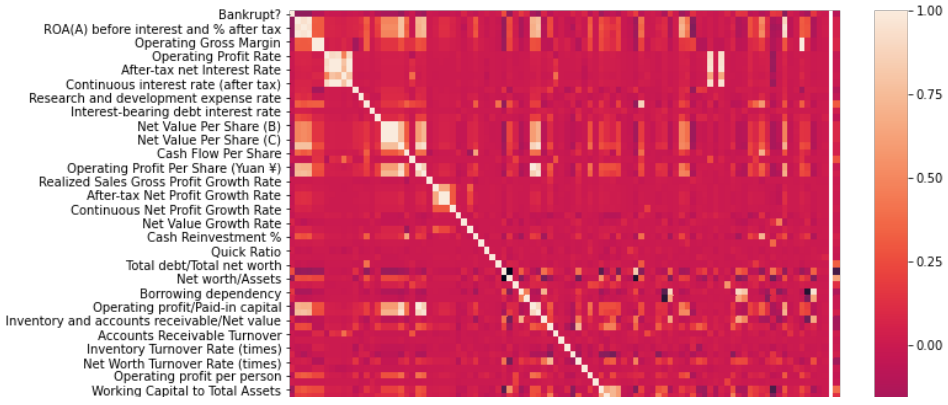
```
# Проверяем категориальные признаки на уникальность
col_obj = data.dtypes[data.dtypes==object].index.values.tolist()
for i in enumerate(col_obj):
    uniq_obj = data[i[1]].unique()
    print(f'{i[0]+1}. {i[1]}: {uniq_obj} | КОЛ-ВО: {len(uniq_obj)}')
```

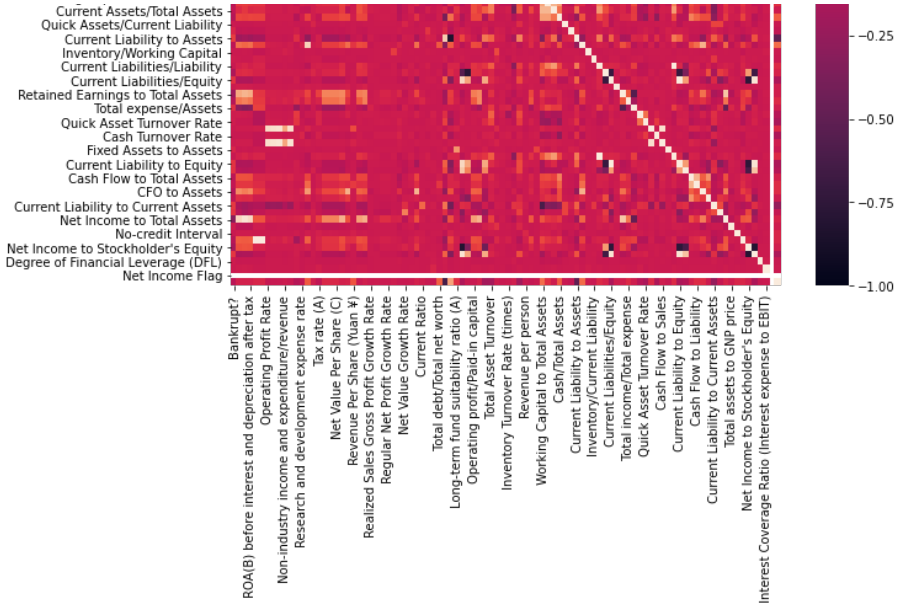
In [10]:

```
# Копируем датасет и применяем label-encoding категориальных признаков для составления корреляционной матрицы
# и последующего применения в модели Random Forest
dataLE = data.copy()
le = LabelEncoder()
col_obj = dataLE.dtypes[dataLE.dtypes==object].index.values.tolist()
for i in col_obj:
    dataLE[i] = le.fit_transform(dataLE[i])
```

In [11]:

```
plt.figure(figsize=(10,10))
g = sns.heatmap(dataLE.corr())
```





In [407]:

```
# Оцениваем важность признаков для целевого
(dataLE.corr()['is_canceled']*100).sort_values(ascending=False)
```

Out[407]:

```
is_canceled                     100.000000
country                          52.533878
arrival_date_year                29.437152
deposit_type                     19.751308
lead_time                         7.588779
market_segment                    5.883349
distribution_channel              4.700574
adults                            4.537695
stays_in_weekend_nights           2.942242
children                          2.469151
stays_in_week_nights              0.049425
reservation_status_date          -0.040024
customer_type                    -0.979502
meal                             -1.987424
reserved_room_type               -2.664975
babies                           -2.954529
agent                            -3.553828
arrival_date_day_of_month        -3.558175
adr                              -4.973463
total_of_special_requests        -8.264548
days_in_waiting_list            -11.344538
arrival_date_month              -16.216285
booking_changes                 -18.118893
assigned_room_type              -19.255699
arrival_date_week_number        -24.489474
required_car_parking_spaces     -29.537194
reservation_status              -87.450209
hotel                                  NaN
is_repeated_guest                      NaN
previous_cancellations                 NaN
previous_bookings_not_canceled         NaN
Name: is_canceled, dtype: float64
```

По результатам корреляционного анализа удаляем столбцы, которые имеют меньшую значимость по отношению к целевому признаку

In [12]:

```
del_data = (dataLE.corr()['Bankrupt?']*100).sort_values(ascending=False)
del_col = del_data[(del_data < 10) & (del_data > -10) | (del_data.isnull())].index.values.tolist()
data.drop(columns=del_col, inplace=True)
dataLE.drop(columns=del_col, inplace=True)
```

In [13]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 32 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Bankrupt?                                               5000 non-null   int64
 1    ROA(C) before interest and depreciation before interest 5000 non-null   float64
 2    ROA(A) before interest and % after tax                  5000 non-null   float64
 3    ROA(B) before interest and depreciation after tax       5000 non-null   float64
 4    Tax rate (A)                                            5000 non-null   float64
 5    Net Value Per Share (B)                                 5000 non-null   float64
 6    Net Value Per Share (A)                                 5000 non-null   float64
 7    Net Value Per Share (C)                                 5000 non-null   float64
 8    Persistent EPS in the Last Four Seasons                 5000 non-null   float64
 9    Operating Profit Per Share (Yuan ¥)                     5000 non-null   float64
 10   Per Share Net profit before tax (Yuan ¥)                5000 non-null   float64
 11   Debt ratio %                                            5000 non-null   float64
 12   Net worth/Assets                                        5000 non-null   float64
 13   Borrowing dependency                                    5000 non-null   float64
 14   Operating profit/Paid-in capital                        5000 non-null   float64
 15   Net profit before tax/Paid-in capital                   5000 non-null   float64
 16   Working Capital to Total Assets                         5000 non-null   float64
 17   Cash/Total Assets                                       5000 non-null   float64
 18   Current Liability to Assets                             5000 non-null   float64
```

```
 19   Working Capital/Equity                                  5000 non-null   float64
 20   Current Liabilities/Equity                              5000 non-null   float64
 21   Retained Earnings to Total Assets                       5000 non-null   float64
 22   Total income/Total expense                              5000 non-null   float64
 23   Total expense/Assets                                    5000 non-null   float64
 24   Current Liability to Equity                             5000 non-null   float64
 25   Equity to Long-term Liability                           5000 non-null   float64
 26   CFO to Assets                                           5000 non-null   float64
 27   Current Liability to Current Assets                     5000 non-null   float64
 28   Liability-Assets Flag                                   5000 non-null   int64
 29   Net Income to Total Assets                              5000 non-null   float64
 30   Net Income to Stockholder's Equity                      5000 non-null   float64
 31   Liability to Equity                                     5000 non-null   float64
dtypes: float64(30), int64(2)
memory usage: 1.2 MB
```

Выполняем **One-hot encoding** для категориальных признаков и масштабирование числовых признаков для применения в **SVM**

In [14]:

```
# Выполняем one-hot encoding и масштабирование для применения в SVM
col_num = data.dtypes[data.dtypes!=object].index.values.tolist()
col_num.remove('Bankrupt?')
se = StandardScaler()
data[col_num] = se.fit_transform(data[col_num])
data = pd.get_dummies(data, drop_first=True)
```

In [15]:

```
TEST_SIZE = 0.3
RANDOM_STATE = 0
```

In [16]:

```
dataLE_X = dataLE.drop(columns='Bankrupt?')
dataLE_y = dataLE['Bankrupt?']
data_X = data.drop(columns='Bankrupt?')
data_y = data['Bankrupt?']
```

In [17]:

```
dataLE_X_train, dataLE_X_test, dataLE_y_train, dataLE_y_test = train_test_split(dataLE_X, dataLE_y, \
                                                                                test_size = TEST_SIZE, \
                                                                                random_state= RANDOM_STATE)
data_X_train, data_X_test, data_y_train, data_y_test = train_test_split(data_X, data_y, \
                                                                        test_size = TEST_SIZE, \
                                                                        random_state= RANDOM_STATE)
```

In [18]:

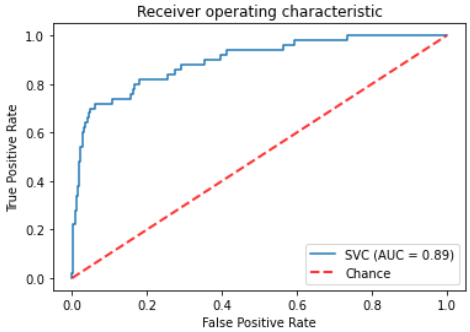
```
def print_metrics(X_train, Y_train, X_test, Y_test, clf):
    clf.fit(X_train, Y_train)
    target = clf.predict(X_test)
    print(f'Сбалансированная оценка: {balanced_accuracy_score(Y_test, target)}')
    fig, ax = plt.subplots()
    plot_roc_curve(clf, X_test, Y_test, ax=ax)
    ax.plot([0, 1], [0, 1], linestyle='--', lw=2, color='r',
            label='Chance', alpha=.8)
    ax.set(xlim=[-0.05, 1.05], ylim=[-0.05, 1.05],
           title="Receiver operating characteristic")
    ax.legend(loc="lower right")
    plt.show()
    print(f'Матрица ошибок:\n {confusion_matrix(Y_test, target)}')
```

1. SVM

In [19]:

```
print_metrics(data_X_train, data_y_train, data_X_test, data_y_test, SVC(random_state=RANDOM_STATE))
```

```
Сбалансированная оценка: 0.5096551724137931
```



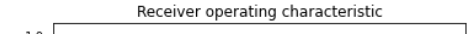

```
Матрица ошибок:
 [[1449    1]
 [  49    1]]
```

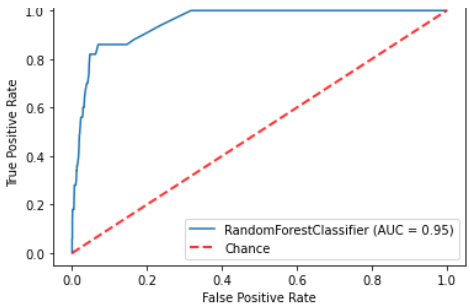
2. Random Forest

In [20]:

```
print_metrics(dataLE_X_train, dataLE_y_train, dataLE_X_test, dataLE_y_test, RandomForestClassifier(random_state=RANDOM_STATE))
```

```
Сбалансированная оценка: 0.5886206896551724
```






```
Матрица ошибок:
 [[1446    4]
 [  41    9]]
```

3. Выводы

В данной работе для оценки моделей были использованы следующие метрики, подходящие для задачи бинарной классификации:

- **balanced accuracy**, так как данная метрика хорошо интерпретируется и используется при несбалансированных классах
- **ROC-кривая (AUC)**, так как позволяет по графику понять, насколько модель может минимизировать **FP (False Positive)**, т.е. признавать отмененным заказ, который таковым не является, и минимизировать **FN (False Negative)**, т.е. признавать бронированным заказ, который был отменен
- **confusion matrix**, так как, хотя и метрикой в полной мере не является, позволяет увидеть общую картину по всем видам ошибок.

По результатам оценивания можно сделать следующий вывод: модель **Random Forest** обладает немного большей предсказательной способностью, чем **Support Vector Machine**. Но при этом обе модели могут использоваться для предсказания, будет ли заказ по бронированию отменен, с минимальным количеством ошибок.

In [ ]: