Малахов В.В. ИУ5Ц-83Б | РК2 - вариант №27

Линейная/логистическая регрессия
Случайный лес
Датасет 27 - https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.cluster import KMeans
from sklearn.cluster import KMeans, AgglomerativeClustering, DBSCAN
from sklearn.metrics import silhouette_score, adjusted_rand_score,
adjusted_mutual_info_score, homogeneity_score
from sklearn.preprocessing import MinMaxScaler,StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier,
GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import learning_curve
from sklearn.model_selection import RandomizedSearchCV
from sklearn.model_selection import RepeatedStratifiedKFold,
StratifiedKFold
import warnings
warnings.filterwarnings('ignore')

df = pd.read_csv('./data.csv')

df.head(3)
```

```
   Bankrupt?   ROA(C) before interest and depreciation before interest
\
0          1                                                 0.370594

1          1                                                 0.464291

2          1                                                 0.426071
```

```
    ROA(A) before interest and % after tax  \
0                                 0.424389
1                                 0.538214
2                                 0.499019

    ROA(B) before interest and depreciation after tax  \
0                                           0.405750
1                                           0.516730
2                                           0.472295

    Operating Gross Margin   Realized Sales Gross Margin  \
0                 0.601457                      0.601457
1                 0.610235                      0.610235
2                 0.601450                      0.601364

    Operating Profit Rate   Pre-tax net Interest Rate  \
0                0.998969                   0.796887
1                0.998946                   0.797380
2                0.998857                   0.796403

    After-tax net Interest Rate   Non-industry income and
expenditure/revenue  \
0                       0.808809
0.302646
1                       0.809301
0.303556
2                       0.808388
0.302035

   ...   Net Income to Total Assets   Total assets to GNP price  \
0  ...                      0.716845                   0.009219
1  ...                      0.795297                   0.008323
2  ...                      0.774670                   0.040003

    No-credit Interval   Gross Profit to Sales  \
0             0.622879               0.601453
1             0.623652               0.610237
2             0.623841               0.601449

    Net Income to Stockholder's Equity   Liability to Equity  \
0                             0.827890             0.290202
1                             0.839969             0.283846
2                             0.836774             0.290189

    Degree of Financial Leverage (DFL)  \
0                             0.026601
1                             0.264577
2                             0.026555

    Interest Coverage Ratio (Interest expense to EBIT)   Net Income
```

```
Flag  \
0                                                 0.564050
1
1                                                 0.570175
1
2                                                 0.563706
1

    Equity to Liability
0              0.016469
1              0.020794
2              0.016474

[3 rows x 96 columns]
```

```
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
```

```
df.describe()
```

```
         Bankrupt?   ROA(C) before interest and depreciation before
interest  \
count  6819.000000                                      6819.000000

mean      0.032263                                         0.505180

std       0.176710                                         0.060686

min       0.000000                                         0.000000

25%       0.000000                                         0.476527

50%       0.000000                                         0.502706

75%       0.000000                                         0.535563

max       1.000000                                         1.000000


        ROA(A) before interest and % after tax  \
count                                6819.000000
mean                                    0.558625
std                                     0.065620
min                                     0.000000
25%                                     0.535543
50%                                     0.559802
75%                                     0.589157
```

```
max                                          1.000000

       ROA(B) before interest and depreciation after tax  \
count                                       6819.000000
mean                                           0.553589
std                                            0.061595
min                                            0.000000
25%                                            0.527277
50%                                            0.552278
75%                                            0.584105
max                                            1.000000

       Operating Gross Margin   Realized Sales Gross Margin  \
count             6819.000000                   6819.000000
mean                 0.607948                      0.607929
std                  0.016934                      0.016916
min                  0.000000                      0.000000
25%                  0.600445                      0.600434
50%                  0.605997                      0.605976
75%                  0.613914                      0.613842
max                  1.000000                      1.000000

       Operating Profit Rate   Pre-tax net Interest Rate  \
count             6819.000000                 6819.000000
mean                 0.998755                    0.797190
std                  0.013010                    0.012869
min                  0.000000                    0.000000
25%                  0.998969                    0.797386
50%                  0.999022                    0.797464
75%                  0.999095                    0.797579
max                  1.000000                    1.000000

       After-tax net Interest Rate  \
count                   6819.000000
mean                       0.809084
std                        0.013601
min                        0.000000
25%                        0.809312
50%                        0.809375
75%                        0.809469
max                        1.000000

       Non-industry income and expenditure/revenue  ...  \
count                                   6819.000000  ...
mean                                       0.303623  ...
std                                        0.011163  ...
min                                        0.000000  ...
25%                                        0.303466  ...
50%                                        0.303525  ...
75%                                        0.303585  ...
```

```
max                                            1.000000  ...

       Net Income to Total Assets   Total assets to GNP price  \
count                 6819.000000               6.819000e+03
mean                     0.807760               1.862942e+07
std                      0.040332               3.764501e+08
min                      0.000000               0.000000e+00
25%                      0.796750               9.036205e-04
50%                      0.810619               2.085213e-03
75%                      0.826455               5.269777e-03
max                      1.000000               9.820000e+09

       No-credit Interval   Gross Profit to Sales  \
count         6819.000000            6819.000000
mean             0.623915               0.607946
std              0.012290               0.016934
min              0.000000               0.000000
25%              0.623636               0.600443
50%              0.623879               0.605998
75%              0.624168               0.613913
max              1.000000               1.000000

       Net Income to Stockholder's Equity   Liability to Equity  \
count                           6819.000000          6819.000000
mean                               0.840402             0.280365
std                                0.014523             0.014463
min                                0.000000             0.000000
25%                                0.840115             0.276944
50%                                0.841179             0.278778
75%                                0.842357             0.281449
max                                1.000000             1.000000

       Degree of Financial Leverage (DFL)  \
count                           6819.000000
mean                               0.027541
std                                0.015668
min                                0.000000
25%                                0.026791
50%                                0.026808
75%                                0.026913
max                                1.000000

       Interest Coverage Ratio (Interest expense to EBIT)   Net
Income Flag  \
count                                          6819.000000
6819.0
mean                                              0.565358
1.0
std                                               0.013214
0.0
```

```
min                                                 0.000000
1.0
25%                                                 0.565158
1.0
50%                                                 0.565252
1.0
75%                                                 0.565725
1.0
max                                                 1.000000
1.0
```

```
        Equity to Liability
count           6819.000000
mean               0.047578
std                0.050014
min                0.000000
25%                0.024477
50%                0.033798
75%                0.052838
max                1.000000

[8 rows x 96 columns]
```

```
# NaN значений нет
df.isna().sum().max()
```

```
0
```

```
# Дубликатов нет
df.duplicated().sum()
```

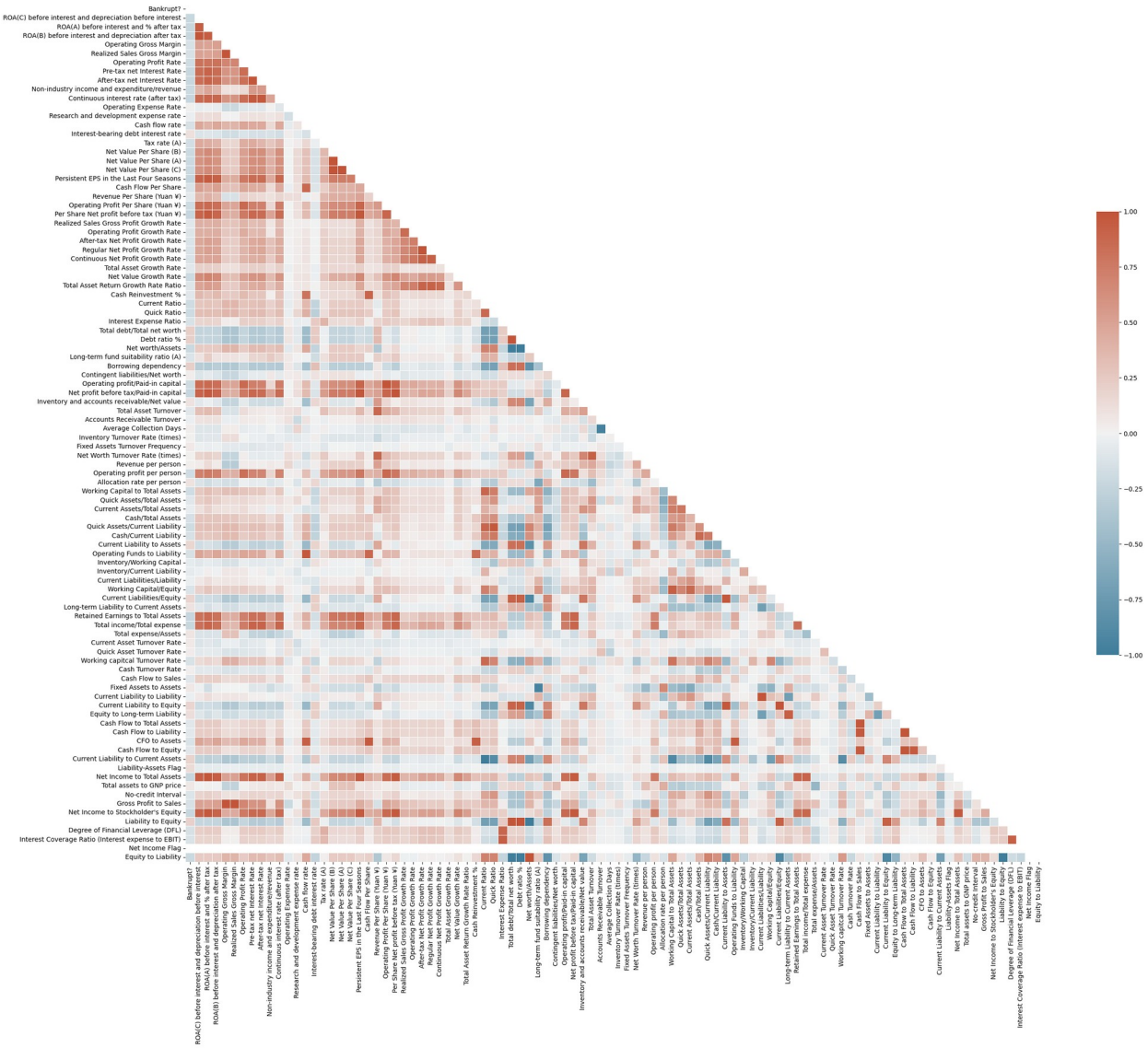
```
0
```

```
# Классы сильно перекошены
print(df['Bankrupt?'].value_counts())
print('-'* 30)
print('Финансово стабильный: ', round(df['Bankrupt?'].value_counts()
[0]/len(df) * 100,2), '%')
print('Финансово нестабильный: ', round(df['Bankrupt?'].value_counts()
[1]/len(df) * 100,2), '%')
```

```
Bankrupt?
0    6599
1     220
Name: count, dtype: int64
------------------------------
Финансово стабильный:  96.77 %
Финансово нестабильный:  3.23 %
```

```
# Матица корреляции по методу Спирмана
```

```
f, ax = plt.subplots(figsize=(30, 25))
mat = df.corr('spearman')
mask = np.triu(np.ones_like(mat, dtype=bool))
cmap = sns.diverging_palette(230, 20, as_cmap=True)
sns.heatmap(mat, mask=mask, cmap=cmap, vmax=1, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5})
plt.show()
```




```
target = "Bankrupt?"
X = df.drop(columns=[target])
y = df[target]

print("X:", X.shape)
print("y:", y.shape)
```

```
X: (6819, 95)
y: (6819,)
```

```
X_train , X_test , y_train , y_test = train_test_split(X , y ,
test_size=0.2)
print("X_train:", X_train.shape)
print("y_train:", y_train.shape)
print("X_test:", X_test.shape)
print("y_test:", y_test.shape)
```

```
X_train: (5455, 95)
y_train: (5455,)
X_test: (1364, 95)
y_test: (1364,)
```

```
# Случайные леса
clf = RandomForestClassifier()
```

```
df_bankrupt_0 = df[df['Bankrupt?'] == 0]
df_bankrupt_1 = df[df['Bankrupt?'] == 1]
```

```
df_bankrupt_0_sample = df_bankrupt_0.sample(n=250, replace=True)
df_bankrupt_1_sample = df_bankrupt_1.sample(n=250, replace=True)
```

```
df_resized = pd.concat([df_bankrupt_0_sample, df_bankrupt_1_sample])
```

```
df_resized.head(3)
```

```
      Bankrupt?   ROA(C) before interest and depreciation before
interest  \
48            0                                          0.493346

5278          0                                          0.581144

1855          0                                          0.506021


       ROA(A) before interest and % after tax  \
48                                   0.550534
5278                                 0.590166
1855                                 0.595235

       ROA(B) before interest and depreciation after tax  \
48                                              0.539804
5278                                            0.611114
1855                                            0.579528

       Operating Gross Margin   Realized Sales Gross Margin  \
48                   0.610206                      0.610206
5278                 0.628944                      0.628944
1855                 0.633210                      0.633210
```

```
      Operating Profit Rate   Pre-tax net Interest Rate  \
48                 0.999023                    0.797429
5278               0.999259                    0.797704
1855               0.999174                    0.797574

      After-tax net Interest Rate  \
48                       0.809341
5278                     0.809530
1855                     0.809564

      Non-industry income and expenditure/revenue  ...  \
48                                       0.303480  ...
5278                                     0.303469  ...
1855                                     0.303418  ...

      Net Income to Total Assets   Total assets to GNP price  \
48                       0.799842                    0.018859
5278                     0.831194                    0.080346
1855                     0.827724                    0.002030

      No-credit Interval   Gross Profit to Sales  \
48              0.623536                0.610202
5278            0.623715                0.628939
1855            0.624111                0.633208

      Net Income to Stockholder's Equity   Liability to Equity  \
48                                0.840599              0.288724
5278                              0.841937              0.275821
1855                              0.841680              0.275506

      Degree of Financial Leverage (DFL)  \
48                                0.028759
5278                              0.026792
1855                              0.026791

      Interest Coverage Ratio (Interest expense to EBIT)   Net Income
Flag  \
48                                             0.568437
1
5278                                           0.565161
1
1855                                           0.565158
1

      Equity to Liability
48                0.017125
5278              0.095179
1855              0.127641
```

```
[3 rows x 96 columns]
```

```
params= {

    "n_estimators":range(25 , 100 , 25),
    "max_depth": range(10 , 70 , 10)

}

model = GridSearchCV(

    clf,
    param_grid= params,
    cv=5,
    n_jobs=-1,
    verbose= 1

)

X = df_resized.drop(columns=['Bankrupt?'])
y = df_resized['Bankrupt?']

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2)

model.fit(X_train , y_train)
model.best_params_
```

```
Fitting 5 folds for each of 18 candidates, totalling 90 fits
```

```
{'max_depth': 40, 'n_estimators': 50}
```

```
y_pred = model.predict(X_test)

# Метрика Accuracy
accuracy = accuracy_score(y_test, y_pred)

# Метрика Precision
precision = precision_score(y_test, y_pred)

# Метрика Recall
recall = recall_score(y_test, y_pred)

# Метрика F1-score
f1 = f1_score(y_test, y_pred)

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```

```
Accuracy: 0.96
Precision: 0.9473684210526315
Recall: 0.9818181818181818
F1-score: 0.9642857142857142

clf = RandomForestClassifier(max_depth=50, n_estimators=75)
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)

Accuracy: 0.96
Precision: 0.9473684210526315
Recall: 0.9818181818181818
F1-score: 0.9642857142857142

df_bankrupt_0 = df[df['Bankrupt?'] == 0]
df_bankrupt_1 = df[df['Bankrupt?'] == 1]

df_bankrupt_0_sample = df_bankrupt_0.sample(n=250, replace=True)
df_bankrupt_1_sample = df_bankrupt_1.sample(n=250, replace=True)

df_resized = pd.concat([df_bankrupt_0_sample, df_bankrupt_1_sample])

X = df_resized.drop(columns=['Bankrupt?'])
y = df_resized['Bankrupt?']

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2)

log_reg = LogisticRegression()

param_grid = {'C': [0.001, 0.01, 0.1, 1, 10, 100],
              'class_weight': ['balanced', None]}

grid_search = GridSearchCV(log_reg, param_grid, cv=5, scoring='f1')

grid_search.fit(X_train, y_train)

best_log_reg = grid_search.best_estimator_

y_pred = best_log_reg.predict(X_test)
```

```
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```

```
Accuracy: 0.64
Precision: 0.6666666666666666
Recall: 0.5306122448979592
F1-score: 0.5909090909090909
```

Вывод: как мы видим, случайные леса при +- равных выборках имеют лучшие показатели метрик, чем логистическая регрессия. Если увеличить размер выборки, то несбалансированность классов также увеличится, что негативно скажется на метриках моделей.

Случайные леса:
Accuracy: 0.96
Precision: 0.9473684210526315
Recall: 0.9818181818181818
F1-score: 0.9642857142857142

Логистическая регрессия:
Accuracy: 0.64
Precision: 0.6666666666666666
Recall: 0.5306122448979592
F1-score: 0.5909090909090909

```
```