(skrzypczak)

Przemysław@lpoz1386 MINGW64 /d/skrzypczak/PARAMETRY

$ python main\_sampling.py

Module 'parameter' started

Initializing parameters ...

Folder checked/created: models/

Folder checked/created: roc\_curves/

Folder checked/created: results

{'random\_seed': 42, 'test\_size': 0.2, 'data\_path\_companies': 'wp11\_companies\_synthetic\_data.parquet', 'data\_path\_links': 'wp11\_b2b\_syntetic\_data\_linked\_1\_only\_rows.parquet', 'balance\_strategy': 'undersample\_pos

itives', 'models\_to\_train': ['K-Nearest Neighbors', 'Naive Bayes', 'Decision Tree', 'Extra Trees', 'Logistic Regression', 'MLP Classifier', 'Quadratic Discriminant Analysis', 'Histogram-Based Gradient Boosting'

, 'XGBoost Classifier', 'LightGBM', 'AdaBoost', 'Random Forest'], 'model\_params': {'K-Nearest Neighbors': {'n\_neighbors': 9, 'weights': 'uniform', 'metric': 'manhattan'}, 'Naive Bayes': {}, 'Decision Tree': {'m

ax\_depth': 3, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'random\_state': 42}, 'Extra Trees': {'max\_depth': 10, 'min\_samples\_split': 2, 'n\_estimators': 50, 'random\_state': 42}, 'Logistic Regression': {'C': 0

.01, 'solver': 'liblinear', 'max\_iter': 1000, 'random\_state': 42}, 'MLP Classifier': {'activation': 'relu', 'alpha': 0.0001, 'hidden\_layer\_sizes': (50,), 'max\_iter': 500, 'random\_state': 42}, 'Quadratic Discrim

inant Analysis': {'reg\_param': 0.0}, 'Histogram-Based Gradient Boosting': {'learning\_rate': 0.1, 'max\_iter': 200, 'max\_depth': None, 'random\_state': 42}, 'XGBoost Classifier': {'learning\_rate': 0.2, 'max\_depth'

: 7, 'n\_estimators': 100, 'subsample': 0.7, 'use\_label\_encoder': False, 'eval\_metric': 'logloss', 'random\_state': 42}, 'LightGBM': {'learning\_rate': 0.1, 'max\_depth': 10, 'n\_estimators': 200, 'num\_leaves': 50,

'random\_state': 42}, 'AdaBoost': {'learning\_rate': 0.01, 'n\_estimators': 50, 'random\_state': 42}, 'Random Forest': {'max\_depth': 5, 'min\_samples\_split': 2, 'n\_estimators': 50, 'random\_state': 42}}, 'models\_fold

er': 'models/', 'roc\_folder': 'roc\_curves/', 'results\_csv\_path': 'training\_results.csv', 'validation\_metrics': ['accuracy', 'precision', 'recall', 'f1-score'], 'detailed\_report': True, 'save\_validation\_report':

True, 'validation\_report\_path': 'results/validation\_report.txt'}

Parameters initialized successfully.

Module 'parameter' done

modul data started

Before balancing:

LINKED

1 206736

0 206736

Name: count, dtype: int64

After balancing:

LINKED

1 206736

0 206736

Name: count, dtype: int64

SUPPLIER BUYER NACE\_SEC\_S NACE\_SEC\_B LINKED

0 7ee008b7ddf6d1e901ca1f33ff1480fb 6c1d59a784145a11da8331d0b8919992 N N 1

1 43dc9ff9c75311756d81410383e42830 6d23560de189d9838678a9c9697dbb9b J N 0

2 0cee8717fc5c5282ef40830d49a9777d 5bbf491ec380e91bcd755ad12b3a14a5 C C 1

3 24c11fa4c0be767870c232e31b5c8d6c ef9f6271393e19f8db5c532fa9e031d5 C C 1

4 3919b880c75d5e6b49618c7044c0e180 f49a4d2dbfdeb79c4c67b72329658362 C C 1

SUPPLIER BUYER NACE\_SEC\_S NACE\_SEC\_B LINKED

413467 e529b6a74dcfe3dd2a8c86dbd1d7291b a9af7d93b1eb4a947f6e26b8a3ed8b40 C N 0

413468 e1223b0cd4921342199cc1aa9f21ece1 d18f6572d555fdd83149f34a9f281085 J M 0

413469 108ce4c4d3cabf5b5d8d27c5da077f92 09bab067e8af4db6195fc7c697dfd039 G C 1

413470 29b23b4510bcc29f01c23ac96044c322 97f978a1ce29f230e14546b663a35227 K F 1

413471 4e2f0197712465c7386c587f005d0536 2e3564dc16901e2a94d31a709df96675 C N 1

number of all rows combined

413472

number of all rows combined loaded as b2b\_df

413472

load data

merge data

SUPPLIER BUYER NACE\_SEC\_S NACE\_SEC\_B LINKED YYYY\_sup DDCCFF\_sup ... ACT\_buyer WAGES\_buyer DIM\_buyer FTA\_buyer FTI\_buyer IAG\_buyer IAI\_buyer

0 7ee008b7ddf6d1e901ca1f33ff1480fb 6c1d59a784145a11da8331d0b8919992 N N 1 2022 110666 ... 5063231.0 0.0 MICRO 445624.0 10400.0 0.0 0.0

1 43dc9ff9c75311756d81410383e42830 6d23560de189d9838678a9c9697dbb9b J N 0 2022 150312 ... 4020246.0 15820970.0 BIG 0.0 0.0 0.0 0.0

2 0cee8717fc5c5282ef40830d49a9777d 5bbf491ec380e91bcd755ad12b3a14a5 C C 1 2022 110666 ... 71890798.0 2916932.0 MEDIUM 767164.0 309486.0 8743.0 3350.0

3 24c11fa4c0be767870c232e31b5c8d6c ef9f6271393e19f8db5c532fa9e031d5 C C 1 2022 111014 ... 115989754.0 11758631.0 BIG 63989831.0 5398081.0 1384274.0 3900.0

4 3919b880c75d5e6b49618c7044c0e180 f49a4d2dbfdeb79c4c67b72329658362 C C 1 2022 110657 ... 27788434.0 591300.0 MEDIUM 4109335.0 660356.0 0.0 0.0

[5 rows x 37 columns]

data validation

The DataFrame does not contain duplicates.

The DataFrame does not contain negative values.

The DataFrame contains missing values (NaN) in the following columns:

CP3\_sup 515

FTA\_sup 12064

FTI\_sup 12064

IAG\_sup 12064

IAI\_sup 12064

CP3\_buyer 611

FTA\_buyer 14060

FTI\_buyer 14060

IAG\_buyer 14060

IAI\_buyer 14060

dtype: int64

DataFrame contains zero values in the following columns:

LINKED 206736

PURCH\_sup 150

TO\_sup 2937

WAGES\_sup 11675

FTA\_sup 15303

FTI\_sup 42585

IAG\_sup 139720

IAI\_sup 222720

PURCH\_buyer 223

TO\_buyer 2405

WAGES\_buyer 10466

FTA\_buyer 16360

FTI\_buyer 41974

IAG\_buyer 142505

IAI\_buyer 222694

dtype: int64

transfer data to training module

modul data done

=== MAIN PIPELINE STARTED ===

[1/4] Loading parameters...

Parameters loaded successfully.

[2/4] Loading data...

Data loaded. Shape: (413472, 37)

[3/4] Training and validating models...

Custom training started on data:

SUPPLIER BUYER NACE\_SEC\_S NACE\_SEC\_B LINKED YYYY\_sup DDCCFF\_sup ... ACT\_buyer WAGES\_buyer DIM\_buyer FTA\_buyer FTI\_buyer IAG\_buyer IAI\_buyer

0 7ee008b7ddf6d1e901ca1f33ff1480fb 6c1d59a784145a11da8331d0b8919992 N N 1 2022 110666 ... 5063231.0 0.0 MICRO 445624.0 10400.0 0.0 0.0

1 43dc9ff9c75311756d81410383e42830 6d23560de189d9838678a9c9697dbb9b J N 0 2022 150312 ... 4020246.0 15820970.0 BIG 0.0 0.0 0.0 0.0

2 0cee8717fc5c5282ef40830d49a9777d 5bbf491ec380e91bcd755ad12b3a14a5 C C 1 2022 110666 ... 71890798.0 2916932.0 MEDIUM 767164.0 309486.0 8743.0 3350.0

3 24c11fa4c0be767870c232e31b5c8d6c ef9f6271393e19f8db5c532fa9e031d5 C C 1 2022 111014 ... 115989754.0 11758631.0 BIG 63989831.0 5398081.0 1384274.0 3900.0

4 3919b880c75d5e6b49618c7044c0e180 f49a4d2dbfdeb79c4c67b72329658362 C C 1 2022 110657 ... 27788434.0 591300.0 MEDIUM 4109335.0 660356.0 0.0 0.0

[5 rows x 37 columns]

Training model: K-Nearest Neighbors

Model 'K-Nearest Neighbors' saved to models/model\_K-Nearest\_Neighbors.pkl

ROC curve saved to roc\_curves/roc\_K-Nearest\_Neighbors.png

Training model: Naive Bayes

Model 'Naive Bayes' saved to models/model\_Naive\_Bayes.pkl

ROC curve saved to roc\_curves/roc\_Naive\_Bayes.png

Training model: Decision Tree

Model 'Decision Tree' saved to models/model\_Decision\_Tree.pkl

ROC curve saved to roc\_curves/roc\_Decision\_Tree.png

Training model: Extra Trees

Model 'Extra Trees' saved to models/model\_Extra\_Trees.pkl

ROC curve saved to roc\_curves/roc\_Extra\_Trees.png

Training model: Logistic Regression

Model 'Logistic Regression' saved to models/model\_Logistic\_Regression.pkl

ROC curve saved to roc\_curves/roc\_Logistic\_Regression.png

Training model: MLP Classifier

Model 'MLP Classifier' saved to models/model\_MLP\_Classifier.pkl

ROC curve saved to roc\_curves/roc\_MLP\_Classifier.png

Training model: Quadratic Discriminant Analysis

Model 'Quadratic Discriminant Analysis' saved to models/model\_Quadratic\_Discriminant\_Analysis.pkl

ROC curve saved to roc\_curves/roc\_Quadratic\_Discriminant\_Analysis.png

Training model: Histogram-Based Gradient Boosting

Model 'Histogram-Based Gradient Boosting' saved to models/model\_Histogram-Based\_Gradient\_Boosting.pkl

ROC curve saved to roc\_curves/roc\_Histogram-Based\_Gradient\_Boosting.png

Training model: XGBoost Classifier

D:\skrzypczak\skrzypczak\Lib\site-packages\xgboost\training.py:183: UserWarning: [11:28:10] WARNING: C:\actions-runner\\_work\xgboost\xgboost\src\learner.cc:738:

Parameters: { "use\_label\_encoder" } are not used.

bst.update(dtrain, iteration=i, fobj=obj)

Model 'XGBoost Classifier' saved to models/model\_XGBoost\_Classifier.pkl

ROC curve saved to roc\_curves/roc\_XGBoost\_Classifier.png

Training model: LightGBM

[LightGBM] [Info] Number of positive: 165434, number of negative: 165343

[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.005549 seconds.

You can set `force\_row\_wise=true` to remove the overhead.

And if memory is not enough, you can set `force\_col\_wise=true`.

[LightGBM] [Info] Total Bins 769

[LightGBM] [Info] Number of data points in the train set: 330777, number of used features: 5

[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500138 -> initscore=0.000550

[LightGBM] [Info] Start training from score 0.000550

Model 'LightGBM' saved to models/model\_LightGBM.pkl

ROC curve saved to roc\_curves/roc\_LightGBM.png

Training model: AdaBoost

Model 'AdaBoost' saved to models/model\_AdaBoost.pkl

ROC curve saved to roc\_curves/roc\_AdaBoost.png

Training model: Random Forest

Model 'Random Forest' saved to models/model\_Random\_Forest.pkl

ROC curve saved to roc\_curves/roc\_Random\_Forest.png

All results saved to training\_results.csv

Custom training done.

Training and validation completed

[4/4] Optional analysis...

Validation results loaded. Total rows: 60

model class precision recall f1-score support value

0 K-Nearest Neighbors 0 0.530735 0.525644 0.528177 41393.0 NaN

1 K-Nearest Neighbors 1 0.529125 0.534211 0.531656 41302.0 NaN

2 K-Nearest Neighbors accuracy NaN NaN NaN NaN 0.529923

3 K-Nearest Neighbors macro avg 0.529930 0.529928 0.529917 82695.0 NaN

4 K-Nearest Neighbors weighted avg 0.529931 0.529923 0.529915 82695.0 NaN

5 Naive Bayes 0 0.545622 0.064576 0.115484 41393.0 NaN

6 Naive Bayes 1 0.502288 0.946104 0.656199 41302.0 NaN

7 Naive Bayes accuracy NaN NaN NaN NaN 0.504855

8 Naive Bayes macro avg 0.523955 0.505340 0.385842 82695.0 NaN

9 Naive Bayes weighted avg 0.523979 0.504855 0.385544 82695.0 NaN

10 Decision Tree 0 0.558146 0.777209 0.649709 41393.0 NaN

11 Decision Tree 1 0.631944 0.383371 0.477230 41302.0 NaN

12 Decision Tree accuracy NaN NaN NaN NaN 0.580507

13 Decision Tree macro avg 0.595045 0.580290 0.563469 82695.0 NaN

14 Decision Tree weighted avg 0.595005 0.580507 0.563564 82695.0 NaN

15 Extra Trees 0 0.559302 0.780736 0.651723 41393.0 NaN

16 Extra Trees 1 0.635707 0.383468 0.478374 41302.0 NaN

17 Extra Trees accuracy NaN NaN NaN NaN 0.582321

18 Extra Trees macro avg 0.597504 0.582102 0.565049 82695.0 NaN

19 Extra Trees weighted avg 0.597462 0.582321 0.565144 82695.0 NaN

20 Logistic Regression 0 0.499720 0.776146 0.607988 41393.0 NaN

21 Logistic Regression 1 0.496550 0.221273 0.306128 41302.0 NaN

22 Logistic Regression accuracy NaN NaN NaN NaN 0.499014

23 Logistic Regression macro avg 0.498135 0.498709 0.457058 82695.0 NaN

24 Logistic Regression weighted avg 0.498137 0.499014 0.457224 82695.0 NaN

25 MLP Classifier 0 0.500762 0.992052 0.665565 41393.0 NaN

26 MLP Classifier 1 0.524566 0.008789 0.017288 41302.0 NaN

27 MLP Classifier accuracy NaN NaN NaN NaN 0.500961

28 MLP Classifier macro avg 0.512664 0.500420 0.341426 82695.0 NaN

29 MLP Classifier weighted avg 0.512651 0.500961 0.341783 82695.0 NaN

30 Quadratic Discriminant Analysis 0 0.582227 0.116493 0.194142 41393.0 NaN

31 Quadratic Discriminant Analysis 1 0.508540 0.916227 0.654055 41302.0 NaN

32 Quadratic Discriminant Analysis accuracy NaN NaN NaN NaN 0.515920

33 Quadratic Discriminant Analysis macro avg 0.545383 0.516360 0.424099 82695.0 NaN

34 Quadratic Discriminant Analysis weighted avg 0.545424 0.515920 0.423846 82695.0 NaN

35 Histogram-Based Gradient Boosting 0 0.576682 0.721378 0.640965 41393.0 NaN

36 Histogram-Based Gradient Boosting 1 0.626957 0.469299 0.536791 41302.0 NaN

37 Histogram-Based Gradient Boosting accuracy NaN NaN NaN NaN 0.595477

38 Histogram-Based Gradient Boosting macro avg 0.601819 0.595339 0.588878 82695.0 NaN

39 Histogram-Based Gradient Boosting weighted avg 0.601792 0.595477 0.588936 82695.0 NaN

40 XGBoost Classifier 0 0.582741 0.685696 0.630040 41393.0 NaN

41 XGBoost Classifier 1 0.617229 0.507942 0.557278 41302.0 NaN

42 XGBoost Classifier accuracy NaN NaN NaN NaN 0.596916

43 XGBoost Classifier macro avg 0.599985 0.596819 0.593659 82695.0 NaN

44 XGBoost Classifier weighted avg 0.599966 0.596916 0.593699 82695.0 NaN

45 LightGBM 0 0.579037 0.714952 0.639856 41393.0 NaN

46 LightGBM 1 0.626448 0.479081 0.542943 41302.0 NaN

47 LightGBM accuracy NaN NaN NaN NaN 0.597146

48 LightGBM macro avg 0.602743 0.597016 0.591400 82695.0 NaN

49 LightGBM weighted avg 0.602717 0.597146 0.591453 82695.0 NaN

50 AdaBoost 0 0.557017 0.778489 0.649389 41393.0 NaN

51 AdaBoost 1 0.630937 0.379522 0.473952 41302.0 NaN

52 AdaBoost accuracy NaN NaN NaN NaN 0.579225

53 AdaBoost macro avg 0.593977 0.579005 0.561670 82695.0 NaN

54 AdaBoost weighted avg 0.593936 0.579225 0.561767 82695.0 NaN

55 Random Forest 0 0.559259 0.779987 0.651433 41393.0 NaN

56 Random Forest 1 0.635209 0.383952 0.478609 41302.0 NaN

57 Random Forest accuracy NaN NaN NaN NaN 0.582188

58 Random Forest macro avg 0.597234 0.581970 0.565021 82695.0 NaN

59 Random Forest weighted avg 0.597192 0.582188 0.565116 82695.0 NaN

=== MAIN PIPELINE DONE ===

(skrzypczak)