Global_Checker_Automation (AE, DC, ZA, MU)

Summary

Training Dataset	3482
Testing Dataset	871
Total Dataset	4353
Unseen dataset	689
Train Test Split	80:20
Model selection	KNeighborsClassifier LogisticRegression RandomForestClassifier DecisionTreeClassifier SVC GradientBoostingClassifier
Model Finalized	LogisticRegression
Countries	AE, DC, MU and ZA
Parameter Used	class_weight, random_state
Add On Stop Words	["kindly", "please","thanks","thank","hi","team", "regard","regards","dear","null","this","de"]
Add on Function	remove_alpha_num() function has been created to remove alpha numeric characters in messages on text cleaning
Word Embeddings	Spacy 3.7.2
SciPy	1.11.3
scikit-learn	1.3.0
python	3.10.0

Hyper Parameter fine tuning	LRparam_grid = {
	'C': [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100],
	'penalty': ['none', 'l1', 'l2', 'elasticnet'],
	'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
	}

Hyper Parameter best_param	{'C': 100, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 100, 'multi_class': 'auto', 'n_jobs': None, 'penalty': 'l2', 'random_state': 149, 'solver': 'liblinear', 'tol': 0.0001, 'verbose': 0, 'warm_start': False}

Data Collection:

- Data has been sourced from the email attachments.
- Sourced data has been collated and synthesised.
- Period of the sourced data is from August to October
- Synthesised data has been splitted based on the data points randomly as Train and Validation set.
- Train data has been used for Train and test split.
- Validation data set are used for unseen data.

Model Data Prep:

- Collected data from the parallel run has been collated using the csv_merge.ipynb script as single file with column heading as "REPORTINGDATE","PARENTTM_ID", "PROCESSID" and "SCSTAR_MESSAGE"
- SCSTAR_MESSAGE column has been taken for the input for the synthetic data for data synthesisation.
- The upcoming steps are done manually.
- Once data has been synthesised the final data column would be in "Date","Tmid","data","GT" as per country wise.
- The country wise data has been collated together and split the data has Train and Validation data set of 80:20.
- Most of the collated data sets are used for Training and minimal data set is used for unseen set.
- The final data set name is Train.csv and Validation.csv.
- The Final attribute of bot Train and validation data set is "Country","Date","ID","Message Details","Forward To"
- X value is Message Details and Y value is Forward To
- "Country", "Date" and "ID" used for identification of each message and GT

Country wise Specification

Train:

Country	Count of Country
AE	160
DC	2077
MU	675
ZA	1441
Grand Total	4353

Validation:

Country	Count of Country
AE	37
DC	259
MU	95
ZA	298
Grand Total	689

Accuracy

Logistic Regression	Train	Test	Validation
Accuracy	98.96%	98.50%	98.54%

Hyper Parameter Tuned Accuracy

Logistic Regression	Train	Test	Validation
Accuracy	99.88%	98.16%	97.96%

Spacy Embeddings:

- spaCy is fast and efficient at runtime, making it a good choice for building production-level NLP applications. One of the essential parts of spaCy is its ability to create and use custom models for specific NLP tasks, such as named entity recognition or part-of-speech tagging.
- spaCy provides 300-dimensional word embeddings for several languages, which have been learned from large corpora. In other words, each word in the model's vocabulary is represented by a list of 300 floating point numbers – a vector – and these vectors are embedded into a 300-dimensional space
- NLTK and spaCy both are very good libraries for building an NLP system. As compared to NLTK, spaCy is more useful in the development and production environment because it provides a very fast and accurate semantic analysis compared to NLTK.
- For the better choice we have used this embedding for our requirement.

Country wise Accuracy Metrics:

Training data

Country	Non-Hyper Tuned			Hyper Tuned	
	Total Dataset	Correctly Predicted	Accuracy	Correctly Predicted	Accuracy
AE	160	158	99%	160	100%
DC	2077	2061	99%	2072	100%
MU	675	646	96%	667	99%
ZA	1441	1439	100%	1441	100%

Average Non-Hyper Tuned: **98%**

Average Hyper Tuned: **100%**

Validation data

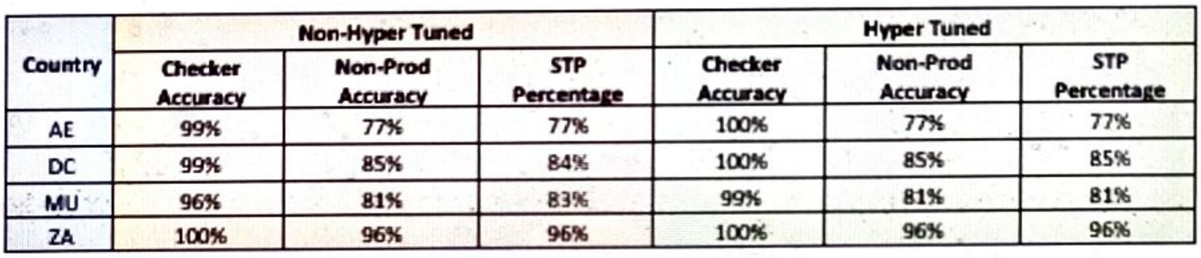
Country	Non-Hyper Tuned			Hyper Tuned	
	Total Dataset	Correctly Predicted	Accuracy	Correctly Predicted	Accuracy
AE	37	37	100%	37	100%
DC	259	257	99%	258	100%
MU	95	93	98%	89	94%
ZA	298	292	98%	291	98%

Average Non-Hyper Tuned: **99%**

Average Hyper Tuned: **98%**

Three-way Recon Metrics:

Training data

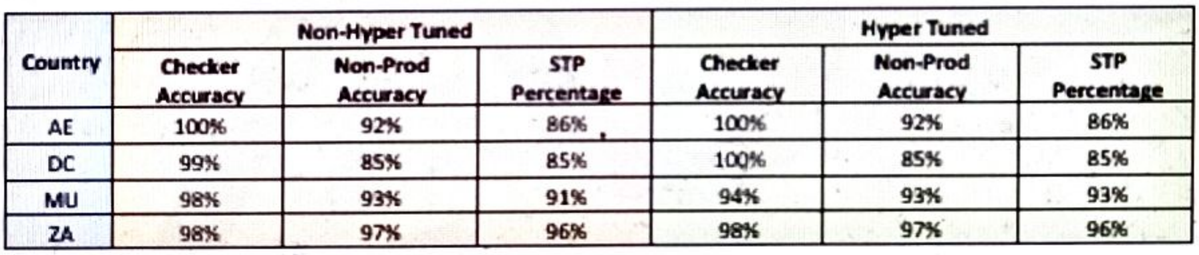


Country	Non-Hyper Tuned			Hyper Tuned		
	Checker Accuracy	Non-Prod Accuracy	STP Percentage	Checker Accuracy	Non-Prod Accuracy	STP Percentage
AE	99%	77%	77%	100%	77%	77%
DC	99%	85%	84%	100%	85%	85%
MU	96%	81%	83%	99%	81%	81%
ZA	100%	96%	96%	100%	96%	96%

Average Non-Hyper Tuned: **85%**

Average Hyper Tuned: **85%**

Validation data



Country	Non-Hyper Tuned			Hyper Tuned		
	Checker Accuracy	Non-Prod Accuracy	STP Percentage	Checker Accuracy	Non-Prod Accuracy	STP Percentage
AE	100%	92%	86%	100%	92%	86%
DC	99%	85%	85%	100%	85%	85%
MU	98%	93%	91%	94%	93%	93%
ZA	98%	97%	96%	98%	97%	96%

Average Non-Hyper Tuned: **90%**

Average Hyper Tuned: **90%**

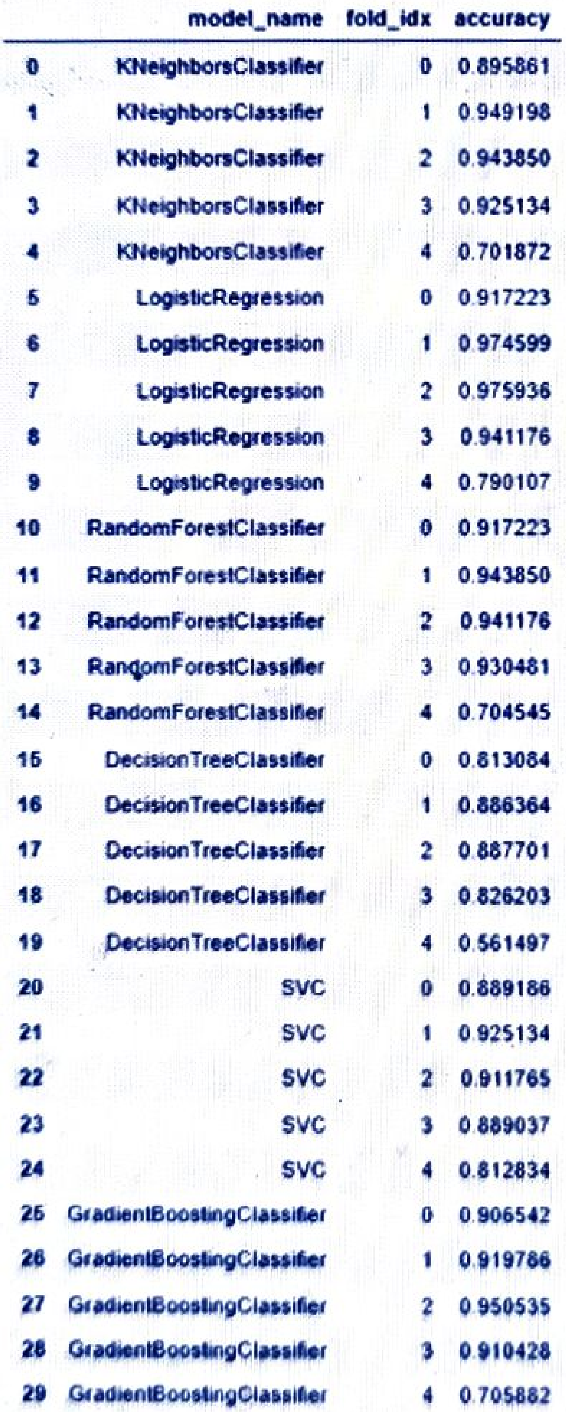
Model Selection:

```
models = [
    KNeighborsClassifier(),
    LogisticRegression(),
    RandomForestClassifier(),
    DecisionTreeClassifier(),
    SVC(),
    GradientBoostingClassifier(),
]
```

Model Selection:

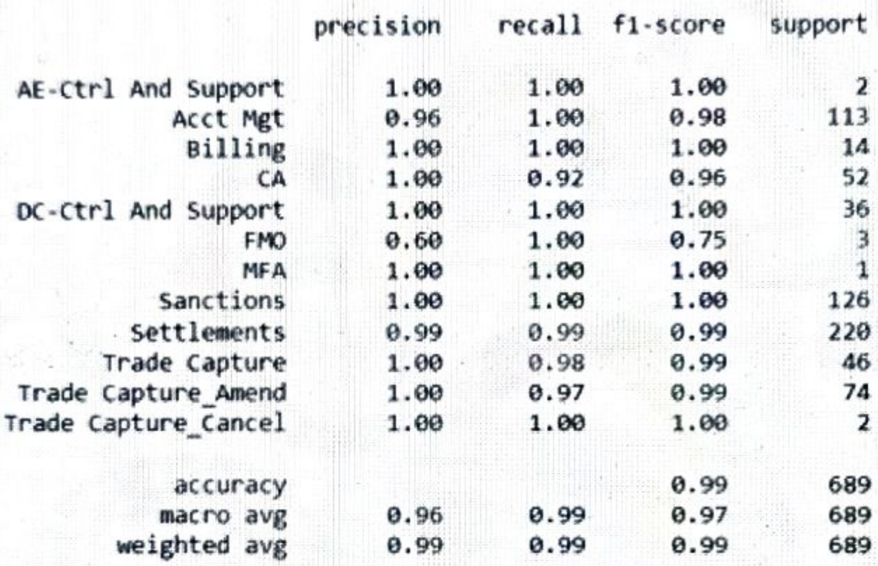
From the below screen shot LogisticRegression performs well.

So currently will experiment with that model



	model_name	fold_idx	accuracy
0	KNeighborsClassifier	0	0.895861
1	KNeighborsClassifier	1	0.949198
2	KNeighborsClassifier	2	0.943850
3	KNeighborsClassifier	3	0.925134
4	KNeighborsClassifier	4	0.701872
5	LogisticRegression	0	0.917223
6	LogisticRegression	1	0.974599
7	LogisticRegression	2	0.975936
8	LogisticRegression	3	0.941176
9	LogisticRegression	4	0.790107
10	RandomForestClassifier	0	0.917223
11	RandomForestClassifier	1	0.943850
12	RandomForestClassifier	2	0.941176
13	RandomForestClassifier	3	0.930481
14	RandomForestClassifier	4	0.704545
15	DecisionTreeClassifier	0	0.813084
16	DecisionTreeClassifier	1	0.886364
17	DecisionTreeClassifier	2	0.887701
18	DecisionTreeClassifier	3	0.826203
19	DecisionTreeClassifier	4	0.561497
20	SVC	0	0.889186
21	SVC	1	0.925134
22	SVC	2	0.911765
23	SVC	3	0.889037
24	SVC	4	0.812834
25	GradientBoostingClassifier	0	0.906542
26	GradientBoostingClassifier	1	0.919786
27	GradientBoostingClassifier	2	0.950535
28	GradientBoostingClassifier	3	0.910428
29	GradientBoostingClassifier	4	0.705882

Validation data



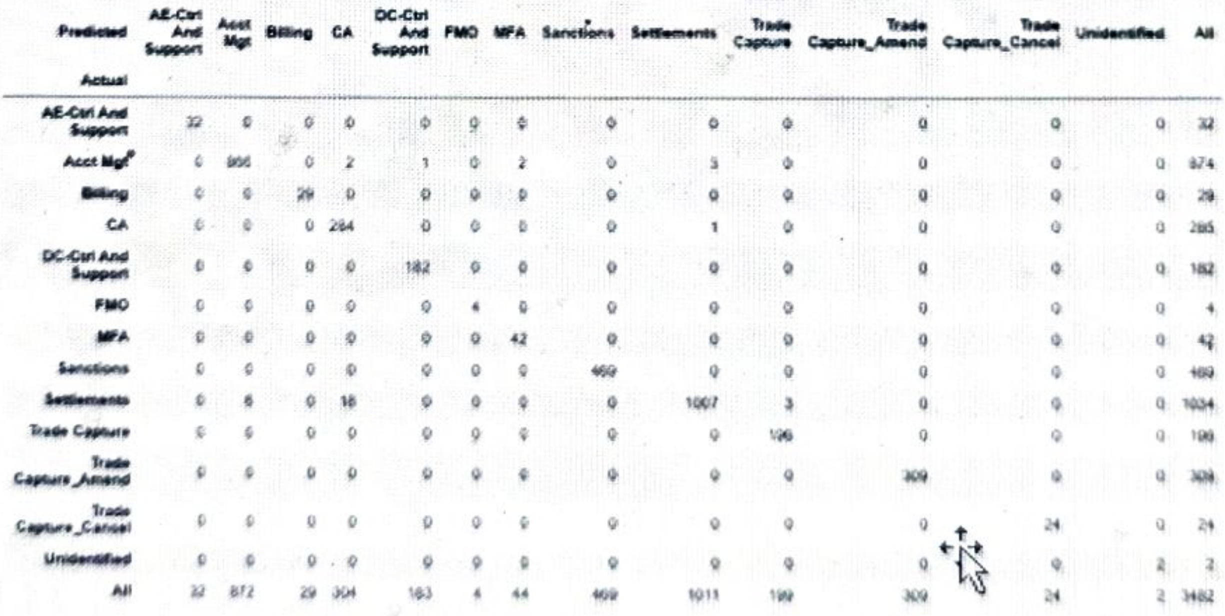
```
                      precision    recall  f1-score   support

 AE-Ctrl And Support       1.00      1.00      1.00         2
            Acct Mgt       0.96      1.00      0.98       113
             Billing       1.00      1.00      1.00        14
                  CA       1.00      0.92      0.96        52
 DC-Ctrl And Support       1.00      1.00      1.00        36
                 FMO       0.60      1.00      0.75         3
                 MFA       1.00      1.00      1.00         1
           Sanctions       1.00      1.00      1.00       126
         Settlements       0.99      0.99      0.99       220
       Trade Capture       1.00      0.98      0.99        46
 Trade Capture_Amend       1.00      0.97      0.99        74
Trade Capture_Cancel       1.00      1.00      1.00         2

            accuracy                           0.99       689
           macro avg       0.96      0.99      0.97       689
        weighted avg       0.99      0.99      0.99       689
```

Confusion Matrix

Training data



Predicted	AE-Ctrl And Support	Acct Mgt	Billing	CA	DC-Ctrl And Support	FMO	MFA	Sanctions	Settlements	Trade Capture	Trade Capture_Amend	Trade Capture_Cancel	Unidentified	All
Actual														
AE-Ctrl And Support	32	0	0	0	0	0	0	0	0	0	0	0	0	32
Acct Mgt	0	866	0	2	1	0	2	0	3	0	0	0	0	874
Billing	0	0	29	0	0	0	0	0	0	0	0	0	0	29
CA	0	0	0	284	0	0	0	0	1	0	0	0	0	285
DC-Ctrl And Support	0	0	0	0	182	0	0	0	0	0	0	0	0	182
FMO	0	0	0	0	0	4	0	0	0	0	0	0	0	4
MFA	0	0	0	0	0	0	42	0	0	0	0	0	0	42
Sanctions	0	0	0	0	0	0	0	469	0	0	0	0	0	469
Settlements	0	6	0	18	0	0	0	0	1007	3	0	0	0	1034
Trade Capture	0	0	0	0	0	0	0	0	0	196	0	0	0	196
Trade Capture_Amend	0	0	0	0	0	0	0	0	0	0	309	0	0	309
Trade Capture_Cancel	0	0	0	0	0	0	0	0	0	0	0	24	0	24
Unidentified	0	0	0	0	0	0	0	0	0	0	0	0	2	2
All	32	872	29	304	183	4	44	469	1011	199	309	24	2	3482

Test data

Predicted	AE-Ctrl And Support	Acct Mgt	Billing	CA	DC-Ctrl And Support	FMO	MFA	Sanctions	Settlements	Trade Capture	Trade Capture_Amend	Trade Capture_Cancel	Unidentified	All
Actual														
AE-Ctrl And Support	9	0	0	0	0	0	0	0	0	0	0	0	0	9
Acct Mgt	0	227	0	1	0	0	0	0	3	0	0	0	0	231
Billing	0	0	11	0	0	0	0	0	0	0	0	0	0	11
CA	0	0	0	70	0	0	0	0	1	0	0	0	0	71
DC-Ctrl And Support	0	0	0	0	51	0	0	1	0	0	0	0	0	52
FMO	0	0	0	0	0	1	0	0	0	0	0	0	0	1
MFA	0	0	0	0	0	0	9	0	0	0	0	0	0	9
Sanctions	0	0	0	0	0	0	0	109	0	0	0	0	0	109
Settlements	0	1	0	4	0	0	0	0	241	0	0	0	0	246
Trade Capture	0	0	0	0	0	0	0	0	0	55	0	0	0	55
Trade Capture_Amend	0	0	0	0	0	0	0	0	0	0	68	0	0	68
Trade Capture_Cancel	0	1	0	0	0	0	0	0	0	0	1	6	0	8
Unidentified	0	0	0	0	0	0	0	0	0	0	0	0	1	1
All	9	229	11	75	51	1	9	110	245	55	69	6	1	871

Validation Data

Predicted	AE-Ctrl And Support	Acct Mgt	Billing	CA	DC-Ctrl And Support	FMO	MFA	Sanctions	Settlements	Trade Capture	Trade Capture_Amend	Trade Capture_Cancel	All
Actual													
AE-Ctrl And Support	2	0	0	0	0	0	0	0	0	0	0	0	2
Acct Mgt	0	113	0	0	0	0	0	0	0	0	0	0	113
Billing	0	0	14	0	0	0	0	0	0	0	0	0	14
CA	0	2	0	48	0	0	0	0	2	0	0	0	52
DC-Ctrl And Support	0	0	0	0	36	0	0	0	0	0	0	0	36
FMO	0	0	0	0	0	3	0	0	0	0	0	0	3
MFA	0	0	0	0	0	0	1	0	0	0	0	0	1
Sanctions	0	0	0	0	0	0	0	126	0	0	0	0	126
Settlements	0	3	0	0	0	0	0	0	217	0	0	0	220
Trade Capture	0	0	0	0	0	0	0	0	1	45	0	0	46
,Trade Capture_Amend	0	0	0	0	0	2	0	0	0	0	72	0	74
Trade Capture_Cancel	0	0	0	0	0	0	0	0	0	0	0	2	2
All	2	118	14	48	36	5	1	126	220	45	72	2	689

ity: Investigate

Accuracy Metrics:

Train and Test data

```
print("Train score:", model.score(X_train, y_train))
print("Test score:", model.score(X_test, y_test))
```

```
Train score: 0.9896611143021252
Test score: 0.9850746268656716
```

Validation data

```
print("Validation score:", log_model.score(x_embeddings, Y))
```

```
Validation score: 0.9854862119013063
```

Hyperparameter Fine Tune:

Param Selection:

```
LRparam_grid = {
    'C': [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100],
    'penalty': ['none', 'l1', 'l2', 'elasticnet'],
    'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
}
```

Grid Search Param:

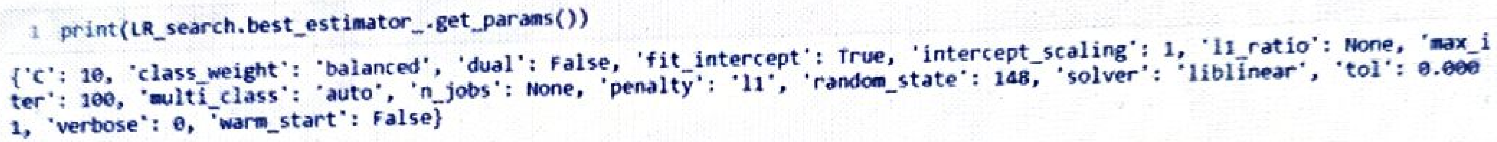
```
LR_search = GridSearchCV(model, param_grid=LRparam_grid, refit = True, verbose = 3, cv=5)
```

Best Param:

```
LR_search.best_params_
```

```
{'C': 10, 'penalty': 'l1', 'solver': 'liblinear'}
```

Best Estimator:

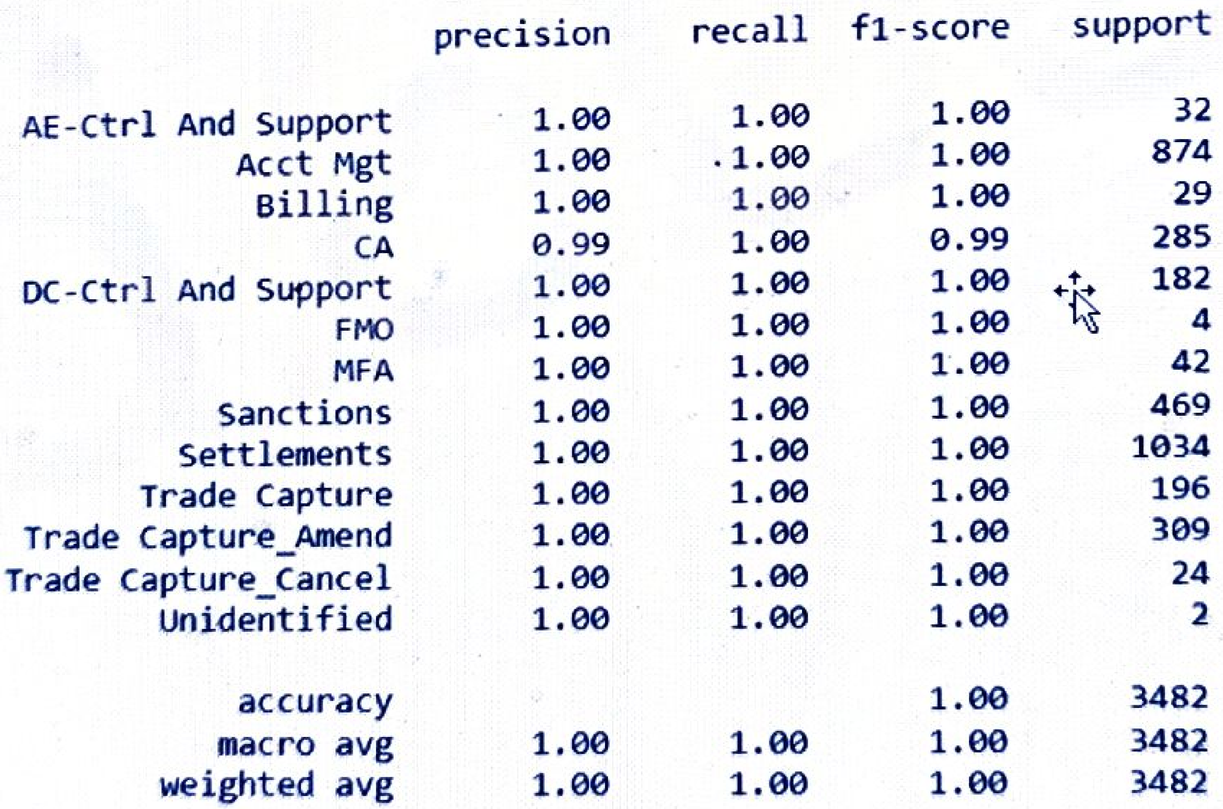


```
print(LR_search.best_estimator_.get_params())
```

```
{'C': 10, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_i
ter': 100, 'multi_class': 'auto', 'n_jobs': None, 'penalty': 'l1', 'random_state': 148, 'solver': 'liblinear', 'tol': 0.000
1, 'verbose': 0, 'warm_start': False}
```

Classification Report:

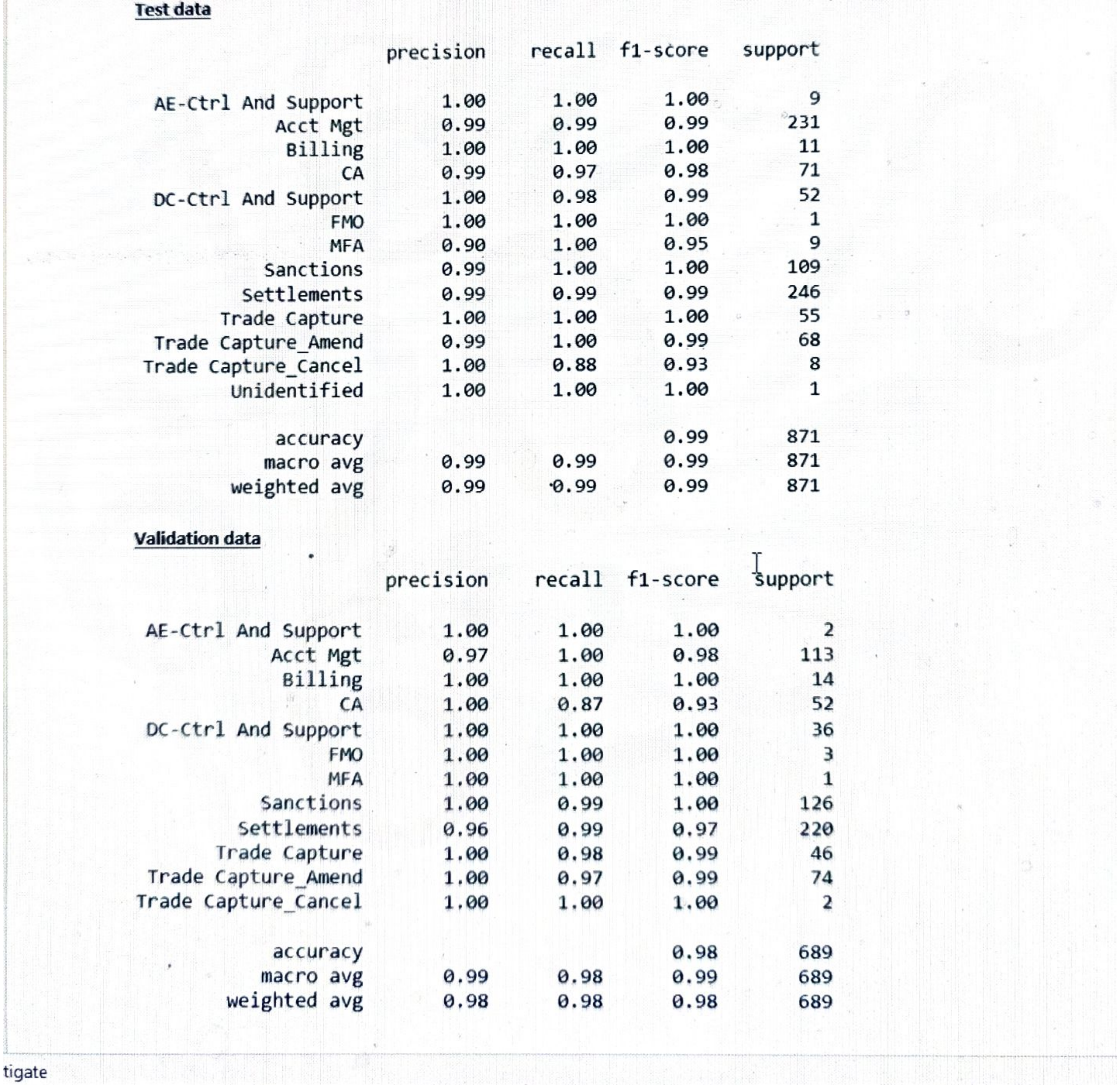
Training data



```
                        precision    recall  f1-score   support

   AE-Ctrl And Support       1.00      1.00      1.00        32
              Acct Mgt       1.00      1.00      1.00       874
               Billing       1.00      1.00      1.00        29
                    CA       0.99      1.00      0.99       285
   DC-Ctrl And Support       1.00      1.00      1.00       182
                   FMO       1.00      1.00      1.00         4
                   MFA       1.00      1.00      1.00        42
             Sanctions       1.00      1.00      1.00       469
           Settlements       1.00      1.00      1.00      1034
         Trade Capture       1.00      1.00      1.00       196
   Trade Capture_Amend       1.00      1.00      1.00       309
  Trade Capture_Cancel       1.00      1.00      1.00        24
          Unidentified       1.00      1.00      1.00         2

              accuracy                           1.00      3482
             macro avg       1.00      1.00      1.00      3482
          weighted avg       1.00      1.00      1.00      3482
```



Test data

```
                        precision    recall  f1-score   support

   AE-Ctrl And Support       1.00      1.00      1.00         9
              Acct Mgt       0.99      0.99      0.99       231
               Billing       1.00      1.00      1.00        11
                    CA       0.99      0.97      0.98        71
   DC-Ctrl And Support       1.00      0.98      0.99        52
                   FMO       1.00      1.00      1.00         1
                   MFA       0.90      1.00      0.95         9
             Sanctions       0.99      1.00      1.00       109
           Settlements       0.99      0.99      0.99       246
         Trade Capture       1.00      1.00      1.00        55
   Trade Capture_Amend       0.99      1.00      0.99        68
  Trade Capture_Cancel       1.00      0.88      0.93         8
          Unidentified       1.00      1.00      1.00         1

              accuracy                           0.99       871
             macro avg       0.99      0.99      0.99       871
          weighted avg       0.99      0.99      0.99       871
```

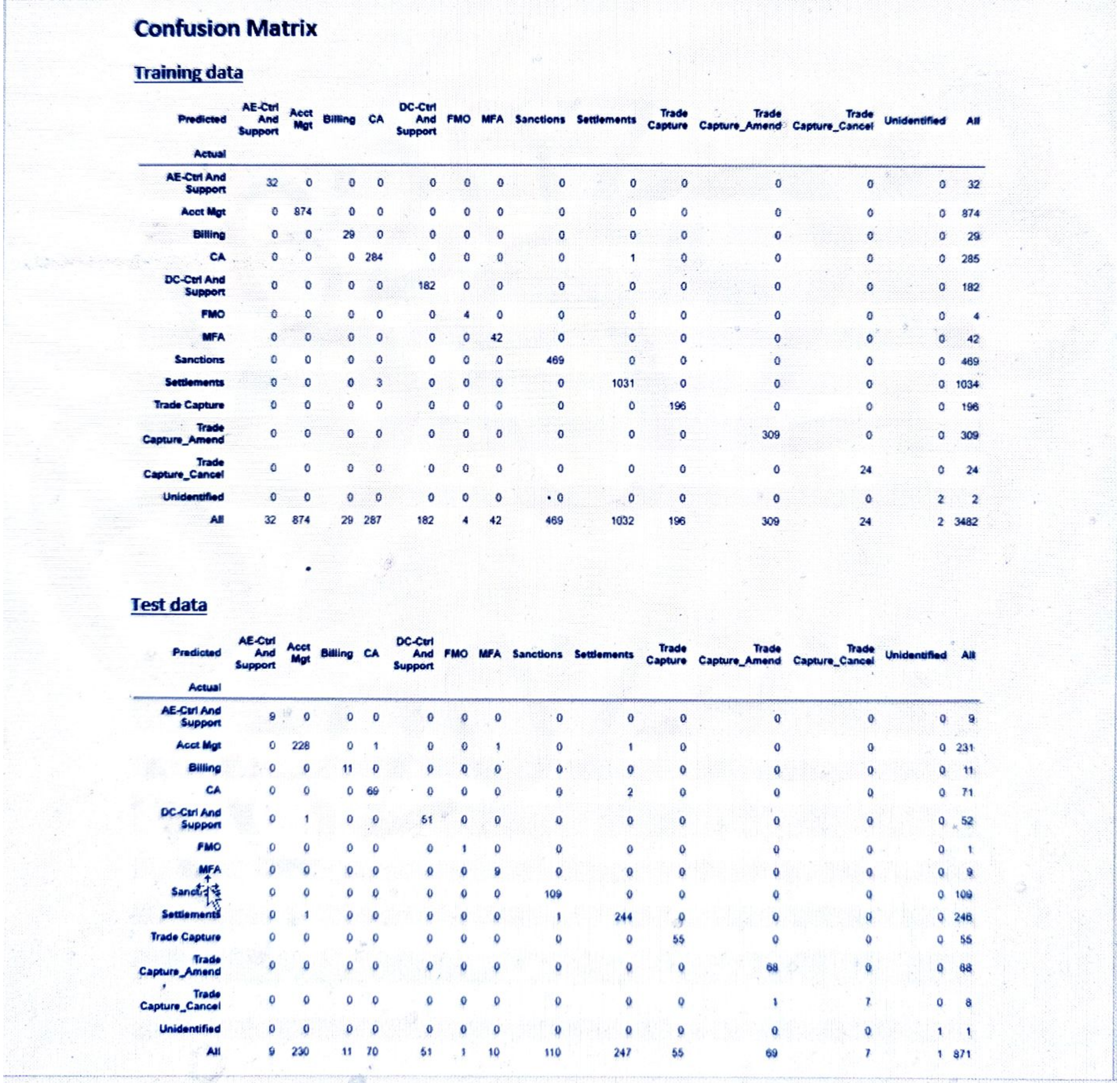
Validation data

```
                        precision    recall  f1-score   support

   AE-Ctrl And Support       1.00      1.00      1.00         2
              Acct Mgt       0.97      1.00      0.98       113
               Billing       1.00      1.00      1.00        14
                    CA       1.00      0.87      0.93        52
   DC-Ctrl And Support       1.00      1.00      1.00        36
                   FMO       1.00      1.00      1.00         3
                   MFA       1.00      1.00      1.00         1
             Sanctions       1.00      0.99      1.00       126
           Settlements       0.96      0.99      0.97       220
         Trade Capture       1.00      0.98      0.99        46
   Trade Capture_Amend       1.00      0.97      0.99        74
  Trade Capture_Cancel       1.00      1.00      1.00         2

              accuracy                           0.98       689
             macro avg       0.99      0.98      0.99       689
          weighted avg       0.98      0.98      0.98       689
```

tigate



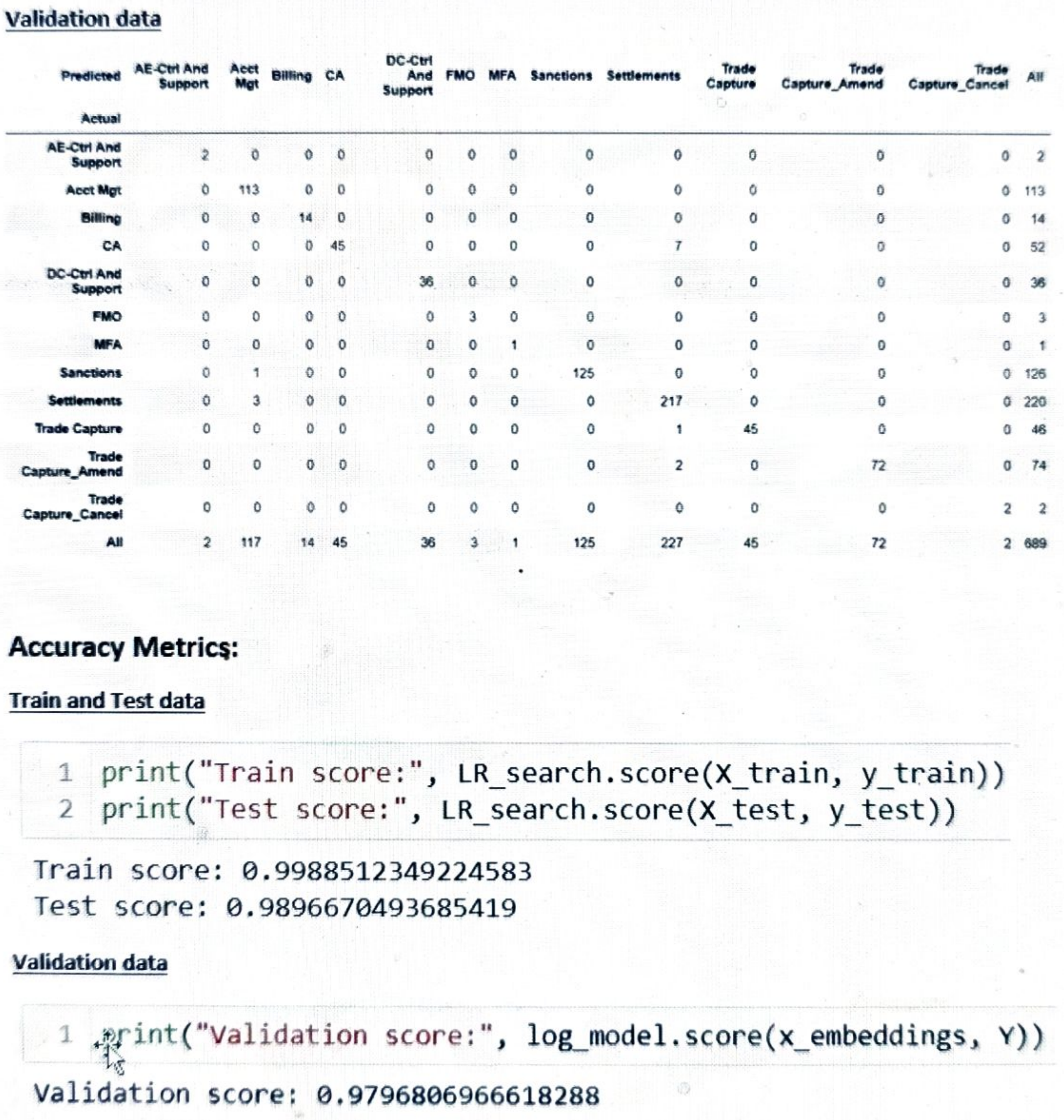
Confusion Matrix

Training data

Predicted Actual	AE-Ctrl And Support	Acct Mgt	Billing	CA	DC-Ctrl And Support	FMO	MFA	Sanctions	Settlements	Trade Capture	Trade Capture_Amend	Trade Capture_Cancel	Unidentified	All
AE-Ctrl And Support	32	0	0	0	0	0	0	0	0	0	0	0	0	32
Acct Mgt	0	874	0	0	0	0	0	0	0	0	0	0	0	874
Billing	0	0	29	0	0	0	0	0	0	0	0	0	0	29
CA	0	0	0	284	0	0	0	0	1	0	0	0	0	285
DC-Ctrl And Support	0	0	0	0	182	0	0	0	0	0	0	0	0	182
FMO	0	0	0	0	0	4	0	0	0	0	0	0	0	4
MFA	0	0	0	0	0	0	42	0	0	0	0	0	0	42
Sanctions	0	0	0	0	0	0	0	469	0	0	0	0	0	469
Settlements	0	0	0	3	0	0	0	0	1031	0	0	0	0	1034
Trade Capture	0	0	0	0	0	0	0	0	0	196	0	0	0	196
Trade Capture_Amend	0	0	0	0	0	0	0	0	0	0	309	0	0	309
Trade Capture_Cancel	0	0	0	0	0	0	0	0	0	0	0	24	0	24
Unidentified	0	0	0	0	0	0	0	0	0	0	0	0	2	2
All	32	874	29	287	182	4	42	469	1032	196	309	24	2	3482

Test data

Predicted Actual	AE-Ctrl And Support	Acct Mgt	Billing	CA	DC-Ctrl And Support	FMO	MFA	Sanctions	Settlements	Trade Capture	Trade Capture_Amend	Trade Capture_Cancel	Unidentified	All
AE-Ctrl And Support	9	0	0	0	0	0	0	0	0	0	0	0	0	9
Acct Mgt	0	228	0	1	0	0	1	0	1	0	0	0	0	231
Billing	0	0	11	0	0	0	0	0	0	0	0	0	0	11
CA	0	0	0	69	0	0	0	0	2	0	0	0	0	71
DC-Ctrl And Support	0	1	0	0	51	0	0	0	0	0	0	0	0	52
FMO	0	0	0	0	0	1	0	0	0	0	0	0	0	1
MFA	0	0	0	0	0	0	9	0	0	0	0	0	0	9
Sanct[illegible]	0	0	0	0	0	0	0	109	0	0	0	0	0	109
Settlements	0	1	0	0	0	0	0	1	244	0	0	0	0	246
Trade Capture	0	0	0	0	0	0	0	0	0	55	0	0	0	55
Trade Capture_Amend	0	0	0	0	0	0	0	0	0	0	68	0	0	68
Trade Capture_Cancel	0	0	0	0	0	0	0	0	0	0	1	7	0	8
Unidentified	0	0	0	0	0	0	0	0	0	0	0	0	1	1
All	9	230	11	70	51	1	10	110	247	55	69	7	1	871



Validation data

Predicted Actual	AE-Ctrl And Support	Acct Mgt	Billing	CA	DC-Ctrl And Support	FMO	MFA	Sanctions	Settlements	Trade Capture	Trade Capture_Amend	Trade Capture_Cancel	All
AE-Ctrl And Support	2	0	0	0	0	0	0	0	0	0	0	0	2
Acct Mgt	0	113	0	0	0	0	0	0	0	0	0	0	113
Billing	0	0	14	0	0	0	0	0	0	0	0	0	14
CA	0	0	0	45	0	0	0	0	7	0	0	0	52
DC-Ctrl And Support	0	0	0	0	36	0	0	0	0	0	0	0	36
FMO	0	0	0	0	0	3	0	0	0	0	0	0	3
MFA	0	0	0	0	0	0	1	0	0	0	0	0	1
Sanctions	0	1	0	0	0	0	0	125	0	0	0	0	126
Settlements	0	3	0	0	0	0	0	0	217	0	0	0	220
Trade Capture	0	0	0	0	0	0	0	0	1	45	0	0	46
Trade Capture_Amend	0	0	0	0	0	0	0	0	2	0	72	0	74
Trade Capture_Cancel	0	0	0	0	0	0	0	0	0	0	0	2	2
All	2	117	14	45	36	3	1	125	227	45	72	2	689

Accuracy Metrics:

Train and Test data

```
print("Train score:", LR_search.score(X_train, y_train))
print("Test score:", LR_search.score(X_test, y_test))
```

```
Train score: 0.9988512349224583
Test score: 0.9896670493685419
```

Validation data

```
print("Validation score:", log_model.score(x_embeddings, Y))
```

```
Validation score: 0.9796806966618288
```