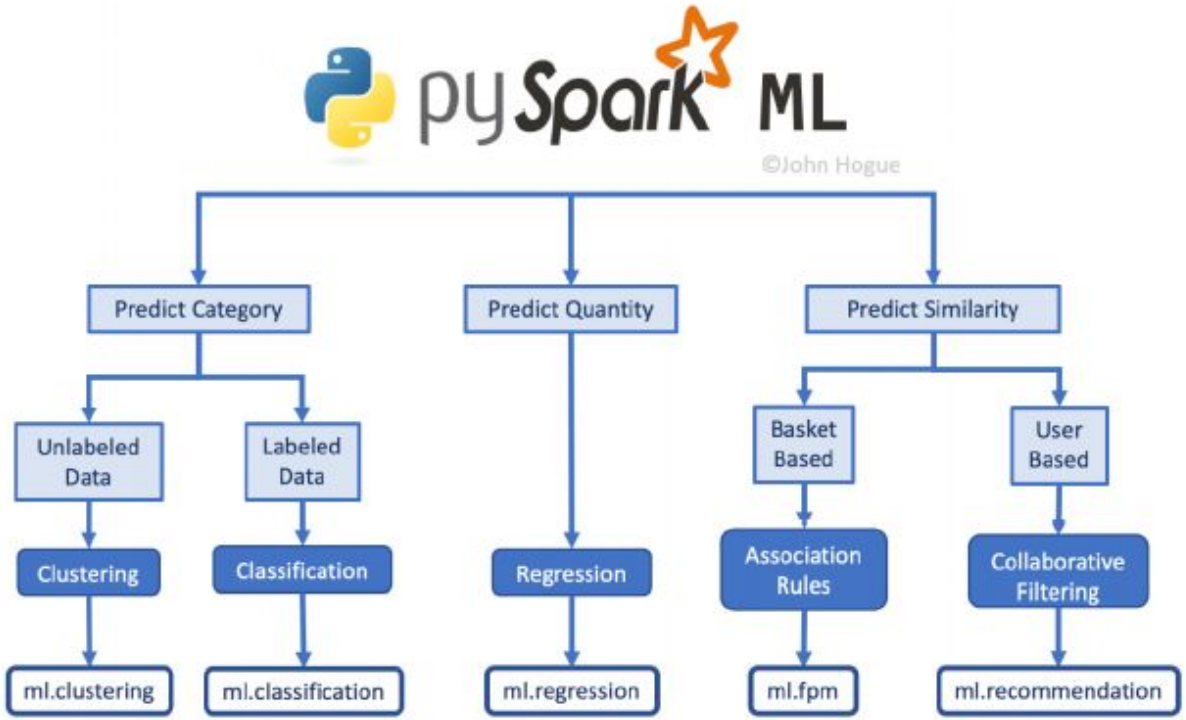
Détection des transactions bancaires frauduleuses à l’aide de la bibliothèque MLLib d’Apache Spark

Présenté par : Lydia MESSAOUI

MLLib de Apache Spark



Données

Description des variables

- **‘type’** type de transaction : payment, transfer, cash out, cash in, debit.
- **‘amount’** montant de la transaction.
- **‘nameOrig’** identifiant de l’émetteur.
- **‘oldbalanceOrg’** solde de l’émetteur avant la transaction.
- **‘newbalanceOrig’** solde de l’émetteur après la transaction.
- **‘nameDest’** identifiant du destinataire.
- **‘oldbalanceDest’** solde du destinataire avant la transaction.
- **‘newbalanceDest’** solde du destinataire après la transaction.
- **‘isFraud’** 1 si la transaction est frauduleuse, sinon 0.
- **‘isFraggedFraud’** 1 si la transaction a été signalée frauduleuse, sinon 0.

EDA : Analyse exploratoire des données

step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
1	TRANSFER	181.0	C1305486145	181.0	0.0	C553264065	0.0	0.0	1	0
1	CASH_OUT	181.0	C840083671	181.0	0.0	C38997010	21182.0	0.0	1	0
1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0

```
root
 |-- step: string (nullable = true)
 |-- type: string (nullable = true)
 |-- amount: string (nullable = true)
 |-- nameOrig: string (nullable = true)
 |-- oldbalanceOrg: string (nullable = true)
 |-- newbalanceOrig: string (nullable = true)
 |-- nameDest: string (nullable = true)
 |-- oldbalanceDest: string (nullable = true)
 |-- newbalanceDest: string (nullable = true)
 |-- isFraud: string (nullable = true)
 |-- isFlaggedFraud: string (nullable = true)
```

```
# Get number of records
print("The data contain %d records." % df.count())
```

```
The data contain 6362620 records.
```

```
# Get number of columns
print("The data contain %d columns." % len(df.columns))
```

```
The data contain 11 columns.
```

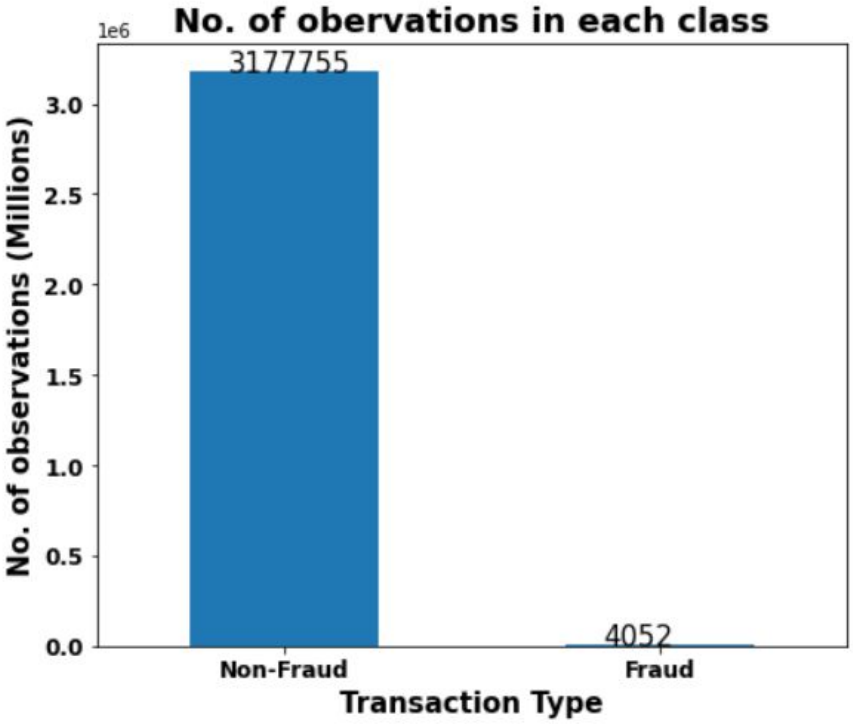
isFraud	count
1	4052
0	3177755

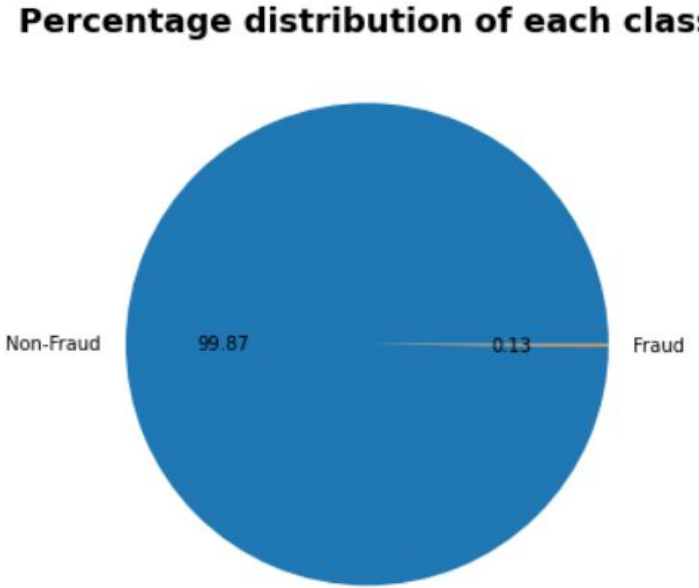
1. Statistique descriptive et visualisations

On peut observer les intervalles [min, max] des chacune des variables, les moyennes et les écart-types.

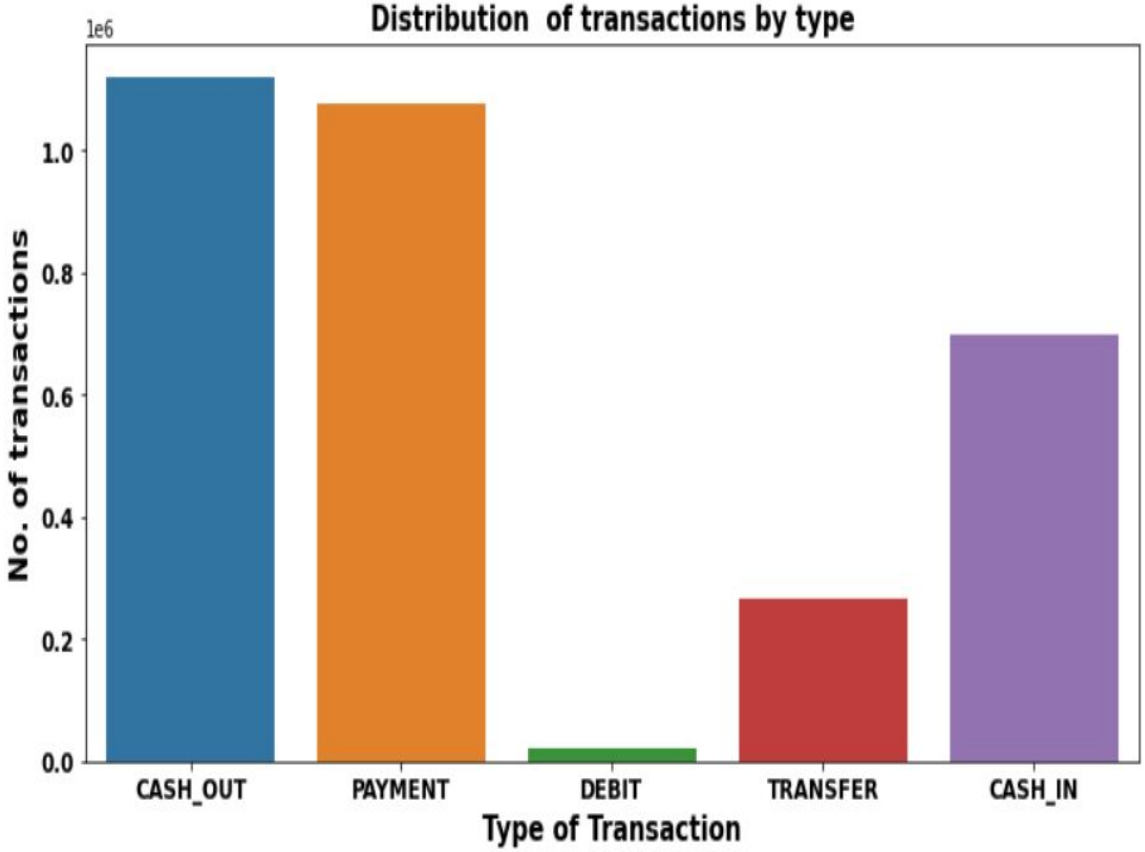
	summary	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud
0	count	3181807	3181807	3181807	3181807	3181807	3181807	3181807
1	mean	243.33075419093615	180376.4924033053	834950.8743586149	856168.4637992503	1101763.094531916	1226409.6227083874	0.0012734901896940952
2	stddev	142.39545491604187	613321.2650007016	2891187.841932541	2927031.285909358	3369890.924303856	3654159.749004548	0.035663269790674586
3	min	1.0	0.0	0.0	0.0	0.0	0.0	0
4	max	743.0	6.98867313E7	5.958504037E7	4.958504037E7	3.5601588935E8	3.5617927892E8	1

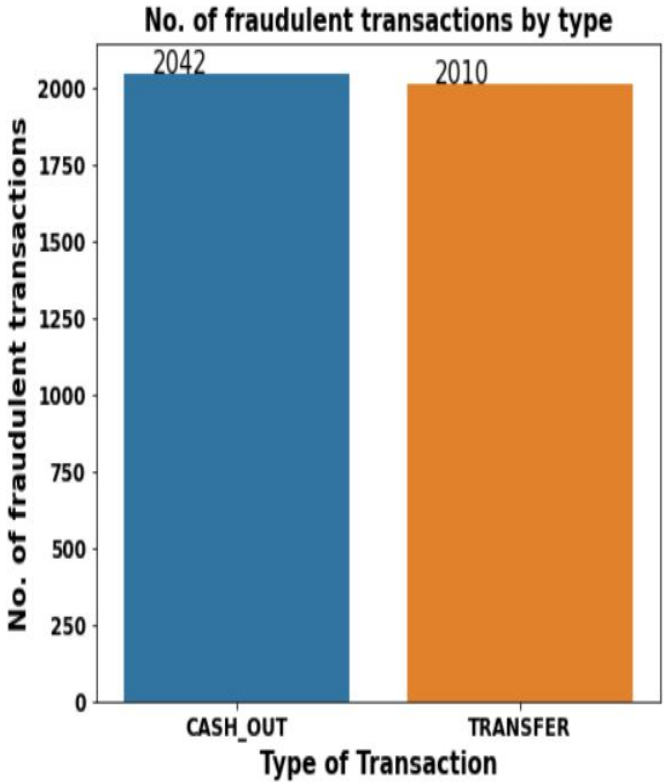
1. Statistique descriptive et visualisations



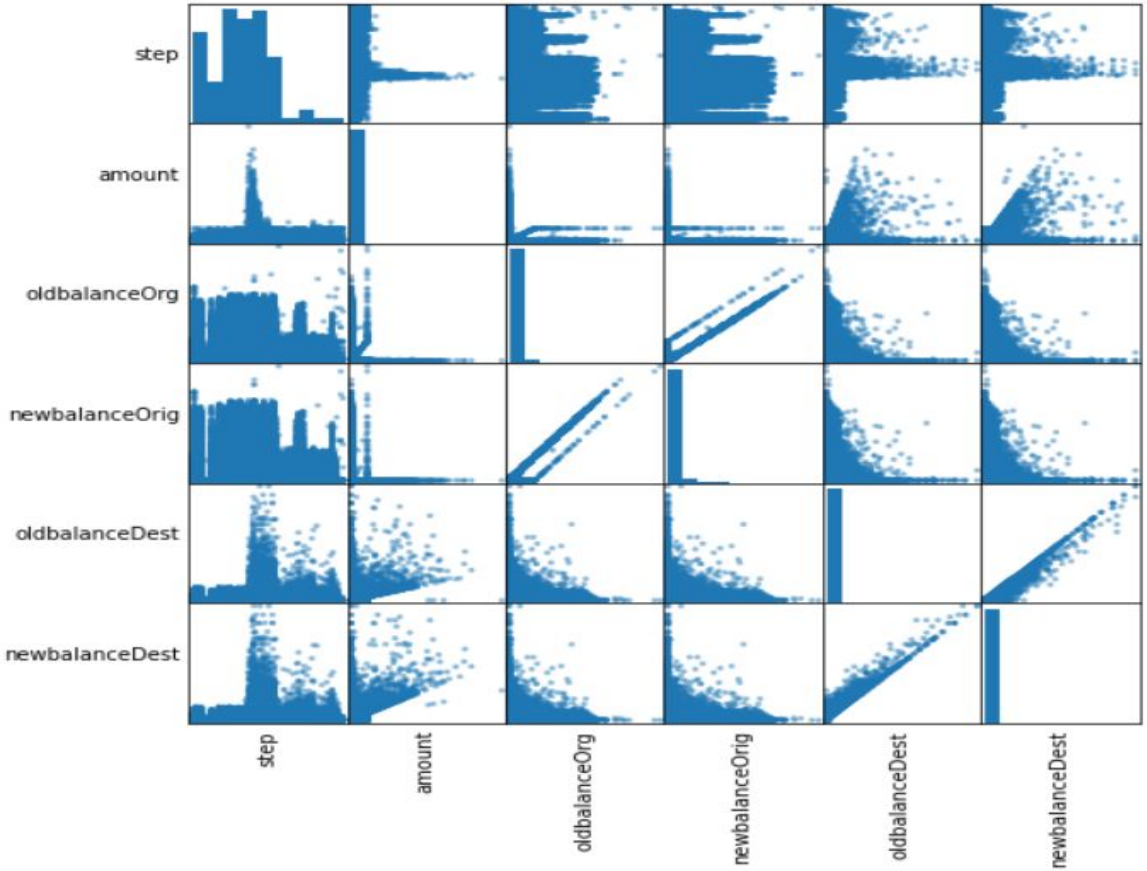


Statistique descriptive et visualisations

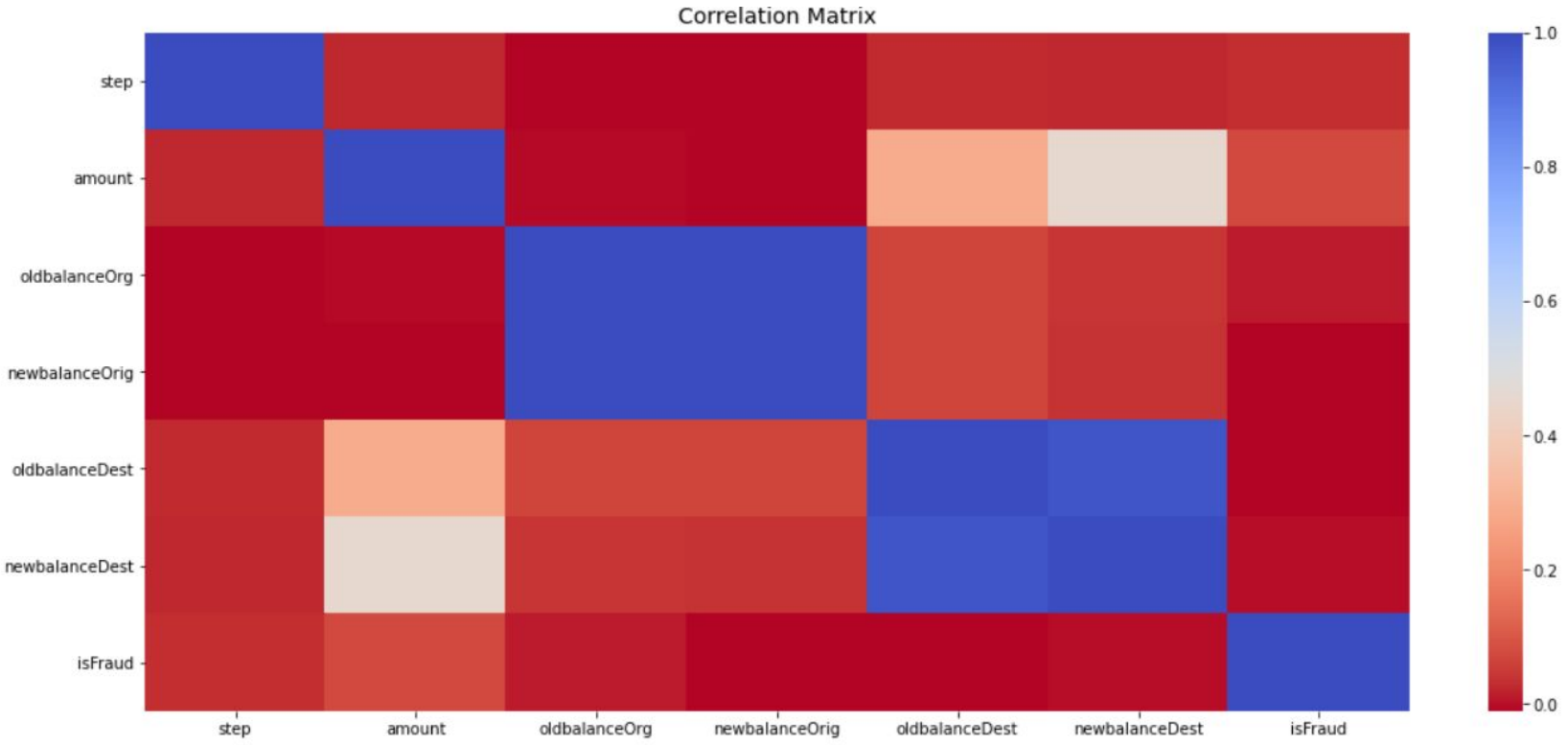




2. Corrélation entre variables



2. Corrélation entre variables



Prétraitement des données

1. Pré-traitements et indexation des données

```
indexed_data.show(5)
```

```
+----+-------+--------------+--------------+-------+-------------+-------------+-----------+--------------+--------------+
|step| amount|newbalanceOrig|oldbalanceDest|isFraud|     num_orig|     num_dest|typeIndexed|nameOrigIndexed|nameDestIndexed|
+----+-------+--------------+--------------+-------+-------------+-------------+-----------+--------------+--------------+
| 1.0|  181.0|           0.0|       21182.0|      1| 8.40083671E8| 8.40083671E8|        0.0|           0.0|           0.0|
| 1.0|7861.64|     168225.59|           0.0|      0|1.912850431E9|1.912850431E9|        1.0|           0.0|           1.0|
| 1.0|9644.94|           0.0|       10845.0|      0|1.900366749E9|1.900366749E9|        4.0|           0.0|           0.0|
| 1.0|2560.74|       2509.26|           0.0|      0|1.648232591E9|1.648232591E9|        1.0|           0.0|           1.0|
| 1.0|1563.82|           0.0|           0.0|      0| 7.61750706E8| 7.61750706E8|        1.0|           0.0|           1.0|
+----+-------+--------------+--------------+-------+-------------+-------------+-----------+--------------+--------------+
only showing top 5 rows
```

2. Transformation des données

```
+--------------------+-------+
|            features|isFraud|
+--------------------+-------+
|(10,[0,1,3,5,6],[...|      1|
|[7861.64,176087.2...|      0|
|[9644.94,4465.0,0...|      0|
|[2560.74,5070.0,2...|      0|
|[1563.82,450.0,0....|      0|
+--------------------+-------+
only showing top 5 rows
```

Classification des transactions

Objectif :

- Effectuer une classification binaire des transactions bancaires

-> comparaison entre les différents modèles de classification

Modèle de classification : Logistic regression

```
lr_model = lr_train(train)
```

```
lr_eval = lr_eval_test(lr_model, test)
```

```
+-------+----------+------+
|isFraud|prediction| count|
+-------+----------+------+
|      1|       0.0|   415|
|      0|       0.0|634751|
|      1|       1.0|   380|
|      0|       1.0|    46|
+-------+----------+------+

Recall :  0.4779874213836478
Precision :  0.892018779342723
F1 Score :  0.6224406224406225
Area under ROC = 0.9909929508193294
Area under PR = 0.5130515281003208
```

Modèle de classification : Decision Tree Classifier

```
dt_eval = Dt_eval_test(dt_model, test)
```

```
+--------------------+---------------+-----------+----------+
|            features|  rawPrediction|probability|prediction|
+--------------------+---------------+-----------+----------+
|[23.31,45360.0,45...|[1437028.0,0.0]|  [1.0,0.0]|       0.0|
|(10,[0,5,6,7,9],[...|[1437028.0,0.0]|  [1.0,0.0]|       0.0|
|[112.56,609035.85...|[1437028.0,0.0]|  [1.0,0.0]|       0.0|
|[154.87,9339.0,91...|[1437028.0,0.0]|  [1.0,0.0]|       0.0|
|[339.82,12076.0,1...|[1437028.0,0.0]|  [1.0,0.0]|       0.0|
+--------------------+---------------+-----------+----------+
only showing top 5 rows

+-------+----------+------+
|isFraud|prediction| count|
+-------+----------+------+
|      1|       0.0|   244|
|      0|       0.0|634761|
|      1|       1.0|   551|
|      0|       1.0|    36|
+-------+----------+------+

Recall :  0.6930817610062893
Precision :  0.938671209540034
F1 Score :  0.7973950795947901
Area under ROC = 0.7798224050687704
Area under PR = 0.4142278715874976
```

Modèle de classification : Random Forest Classifier

```
rf_model = rf_train(train)
```

```
rf_eval = rf_eval_test(rf_model, test)
```

```
+-------+----------+------+
|isFraud|prediction| count|
+-------+----------+------+
|      1|       0.0|   479|
|      0|       0.0|634796|
|      1|       1.0|   316|
|      0|       1.0|     1|
+-------+----------+------+

Recall :  0.39748427672955977
Precision :  0.9968454258675079
F1 Score :  0.5683453237410073
Area under ROC = 0.9692023953420934
Area under PR = 0.7132785942938621
```

Modèle de classification : Gradient-Boosted Tree Classifier

```
GBT_model = GBT_train(train)
```

```
GBT_eval = GBT_eval_test(GBT_model, test)
```

```
+-------+----------+------+
|isFraud|prediction| count|
+-------+----------+------+
|      1|       0.0|   237|
|      0|       0.0|634783|
|      1|       1.0|   558|
|      0|       1.0|    14|
+-------+----------+------+

Recall :  0.7018867924528301
Precision :  0.9755244755244755
F1 Score :  0.8163862472567667
Area under ROC = 0.9829058015208803
Area under PR = 0.7807439732858391
```

Modèle de classification : Naive Bayes

```
NB_model = NB_train(train)
```

```
NB_eval = NB_eval_test(NB_model, test)
```

```
+-------+----------+------+
|isFraud|prediction| count|
+-------+----------+------+
|      1|       0.0|   355|
|      0|       0.0|596227|
|      1|       1.0|   440|
|      0|       1.0| 38570|
+-------+----------+------+

Recall :  0.5534591194968553
Precision :  0.011279159189951295
F1 Score :  0.022107775405099866
Area under ROC = 0.49412195289727784
Area under PR = 0.0012358315240476004
```

Classification des transactions

Observation: Le GRadient Boost Classifier est plus performant que les autres modèles si on considère toutes les métriques de l'évaluation. Donc c'est celui qui sera utilisé pour classer les nouvelles transactions.

FIN