Prédire une transaction financière frauduleuse via mobile money

Regis M OBIANG MBA

05/06/2020

1. Contexte

Plusieurs pays africains et asiatiques ont été limités face aux problèmes de faibles bancarisations de leurs populations du à la faible représentativité de ces banques dans la géographie du pays, souvent concentrer dans les grandes capitales. Pour répondre à cette problématique, depuis plusieurs décennies des pays africains et asiatiques ont déployés le mobile money. Mobile money est un porte-monnaie électronique lié essentiellement à un opérateur de téléphonie mobile, et donc certains opérateurs téléphonie mobiles ont des partenariats avec les banques constitutionnelles. Les différents flux d'un service de mobile money peut être schématisé de façon simpliste comme suit :

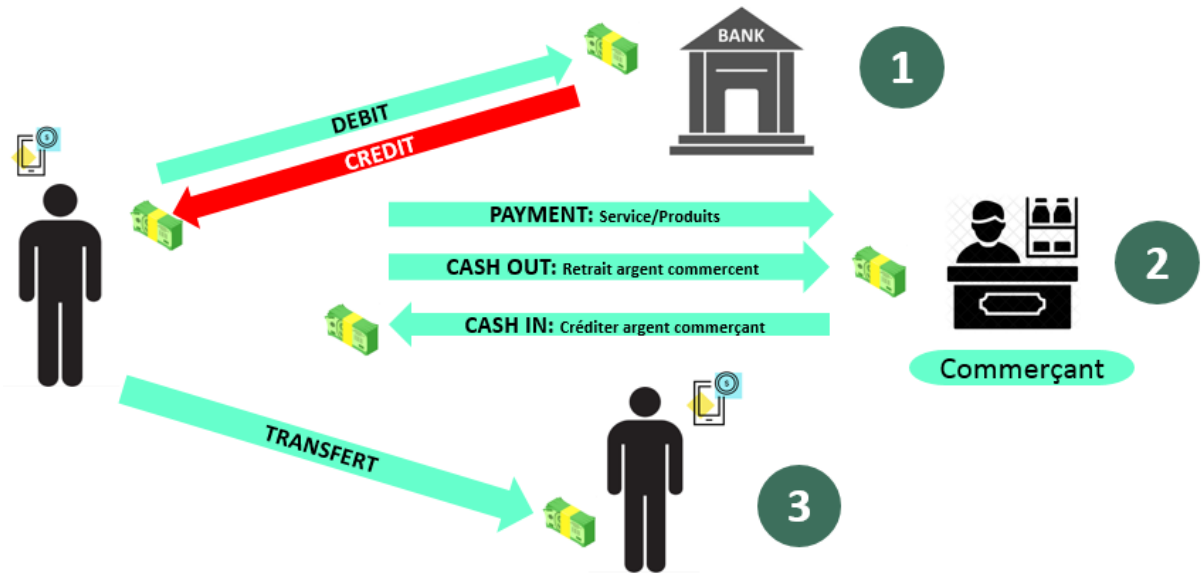

Fig. 1 : Schéma des flux d'un service de mobile money

L'objectif du projet a consisté à prédire une transaction financière frauduleuse qui pourrait survenir lors d'un service de mobile money.

2. Méthode

2.1. Environnement

Le traitement des données, l'analyse exploratoire et la modélisation du Machine Learning (ML) ont été réalisés sous Python et certaines visualisations sous Tableau.

2.2. Gestion des données

Les données sont issues d'un simulateur de transactions des services de paiements mobiles postées en ligne sur le site de Kaggle (https://www.kaggle.com/ntnu-testimon/paysim1).

i. Structure des données

Le fichier de données csv est composé de près 6,4 millions observations avec 11 variables. Les sept variables clés sont :

Nom de la variable	Description	Type
isFraud (outcome)	Transaction frauduleuse (Oui/Non)	Qualitative
type	Type de transaction (Debit, Payment, Cash in, Cash out et Transfert)	Qualitative
amount	Montant du transaction	Quantitative
oldbalanceOrg	Solde initial avant la transaction du destinataire	Quantitative
newbalanceOrg	Solde après la transaction du destinataire	Quantitative
oldbalanceDest	Solde initial avant la transaction du destinateur	Quantitative
newbalanceDest	Solde après la transaction du destinateur	Quantitative

ii. Données pour le modèle ML

La structure des données a été considérée de deux façons, premièrement une base de données avec toutes les sept variables et deuxièmement une base de données des variables présélectionnés préalablement via une régression logistique multivariée en retenant seulement les variables statistiquement significatives avec la variable d'intérêt (isFraud). Les deux bases ont été respectivement subdivisées au ratio 80:20 pour les données d'apprentissage et les données tests.

2.3. Analyse exploratoire

Les variables quantitatives ont été décrites avec les mesures de tendances centrales telles que la moyenne et la médiane, et les histogrammes pour vérifier la distribution de gaussienne. Les variables quantitatives asymétriques ont été log transformées. Les variables qualitatives ont été décrites avec les tables de fréquences et des diagrammes en barres. Les associations entre la variable d'intérêt (isFraud) et les features ont été décrites via les box plots (pour les variables quantitatives) et bubble plots (pour les variables qualitatives).

2.4. Modélisation Machine Learning

I. Choix du modèle ML

Pour répondre à la problématique de prédiction de transaction frauduleuse, le modèle de régression logistique a été choisi comme modèle de base. De plus, pour prendre en compte le contexte asymétriques des données quantitatives et le déséquilibre de la variable d'intérêt (isFraud), le modèle de décision trees a été choisi pour sa robustesse.

II. Mesures d'évaluation du modèle ML

Les différents modèles ML ont été évalués en utilisant les mesures Recall et Precision

Recall	VP/(VP+FN)	Capacité du modèle ML a classifié une transaction frauduleuse, si elle est frauduleuse
Precision	VP/(VP+FP)	Capacité du modèle ML a classifié correctement une transaction frauduleuse
VP : Vrai Positif ; FP : Faux Positif ; FN : Faux Négatif		

3. Résultats

I. Analyse exploratoire

L'analyse exploratoire des données a montré que moins de 1% (soit 8.213/6.362.620) de toutes les transactions étaient des transactions frauduleuses. Plus de 91% des transactions (soit 5.788.279/6.362.620) ont été type (Cash out (n = 2.237.500), Payment (n = 2.151.495) et Cas in (n = 1.399.284)). La tendance centrale (moyenne et médiane) des variables quantitatives a montré des grands écarts, ainsi l'observation l'asymétrique des données quantitatives qui a été partiellement corrigée par une transformation logarithme.

L'association entre le type de transaction et la variable d'intérêt (isFraud) semble montrer que les transactions du type « Cash out » et « Transfert » sont susceptibles à des transactions financières frauduleuses. De plus, il a été observé que les montants de transactions et le solde initial avant la transaction du destinataire élevés ont semblé être associés à une transaction financière frauduleuse.

II. Machine Learning :

- Régression Logistique :

Mesures	Toutes variables	Variables présélectionnées
Train score	**99%**	**99%**
Test score	**99%**	**99%**
Recall	**89%**	**88%**
Precision	**53%**	**51%**

- Decision trees :

Mesures	Toutes variables	Variables présélectionnées
Train score	**99%**	**99%**
Test score	**99%**	**99%**
Recall	**92%**	**89%**
Precision	**89%**	**90%**

4. Conclusions

Les modèles ML de decision trees ont semblé donner les meilleurs résultats comparativement aux modèles de régression logistique. Toutefois, les scores du train et test sont restés excessivement élevés, ce qui pourrait cacher un sur-apprentissage des différents modèles de ML.

Des améliorations dans l'avenir qui pourraient être d'envisager un approfondissement les différentes causes du sur-apprentissage du modèle ML. De plus, d'appliquer un modèle ML plus robuste XgBoost afin de prendre en compte l'asymétrique des données quantitatives et le déséquilibre des groupes de la variable d'intérêt.