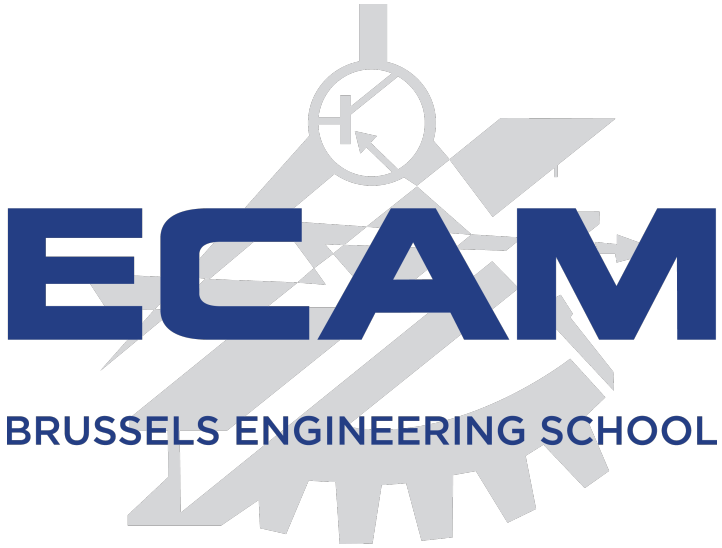
Rapport de laboratoire:
Intelligence Artificielle

Matthias Léonard et Dawid Krasowski

2022/2023



Superviseur : Monsieur HASSELMANN Ken

Contents

1 Introduction **3**

2 Présentation des données **3**
2.1 Description des données 3
2.2 identification des Features 3

3 Conclusion **5**

1 Introduction

Dans le cadre des laboratoires d'intelligence artificielle dispensé à l'ECAM en 2ème Master en ingénieur informatique. Sous la supervision de Monsieur HASSELMANN notre projet se base sur la recherche de fraude à la carte bancaire. Nous travaillons sur un jeu de données de transactions bancaires issues d'Ouganda ces données on été fournis lors d'une compétition Xente de 2019.
`https://zindi.africa/competitions/xente-fraud-detection-challenge`

Notre projet est disponible sur GitHub à l'adresse suivante :
`https://github.com/LeTouristeDeLECAM/Lab_AI_Fraud_Detection`
Pour des questions de propriété et droit les données ne sont pas disponibles sur GitHub.

2 Présentation des données

2.1 Description des données

Column Name	Definition	Type
TransactionId	Unique transaction identifier on platform	object
BatchId	Unique number assigned to a batch of transactions for processing	object
AccountId	Unique number identifying the customer on platform	object
SubscriptionId	Unique number identifying the customer subscription	object
CustomerId	Unique identifier attached to Account	object
CurrencyCode	Country currency	object
CountryCode	Numerical geographical code of country	int64
ProviderId	Source provider of Item bought.	object
ProductId	Item name being bought.	object
ProductCategory	ProductIds are organized into these broader product categories.	object
ChannelId	"Identifies if customer used web;Android; IOS; pay later or checkout."	object
Amount	Value of the transaction. Positive for debits from customer account and negative for credit into customer account	float64
Value	Absolute value of the amount	int64
TransactionStartTime	Transaction start time	object
PricingStrategy	Category of Xente's pricing structure for merchants	int64
FraudResult	Fraud status of transaction 1 -yes or 0-No	int64

Table 1: Description des données

2.2 identification des Features

Dans un premier temps nous cherchons à identifier les features qui vont nous permettre de créer un modèle de prédiction de fraude.

Nous pouvons observer que certaines données ne sont pas utile pour obtenir un modèle. Nous décidons de supprimer les colonnes suivantes :

- CurrencyCode : Toutes les transactions sont en UGX soit en Shilling Ougandais.
- CountryCode : Toutes les transactions sont en Ouganda.

Nous pouvons également imaginer à première vue que les données Amount et Value sont similaires. Néanmoins suite à une analyse:

diff = test2["Amount"] - abs(test2["Value"])
diff.describe()

nous pouvons observer que les deux colonnes ne sont pas totalement identiques.

Methods	Value
count	95662.0
mean	-3182.7375081014397
std	17692.308422485323
min	-2000000.0
25%	-100.0
50%	0.0
75%	0.0
max	0.0

Table 2: Description statistique de la différence entre Amount et Value

Nous avons décidé de garder les données Amount et Value. Car sur les 193 fraudes que comporte le jeu de données, 17 fraudes sont réalisé quand Amount et Value sont différents (8,8%).

Features à supprimer : Nous pouvons observer que productCategory et productID sont fortement corrélés. il en est de même pour amount et value.

Pour pousuivre notre analyse nous réalisons une analyse en composante principale (ACP) sur les données.
Cette analyse nous permet de réduire la dimensionnalité de nos données et identifier les données qui sont les plus importantes.

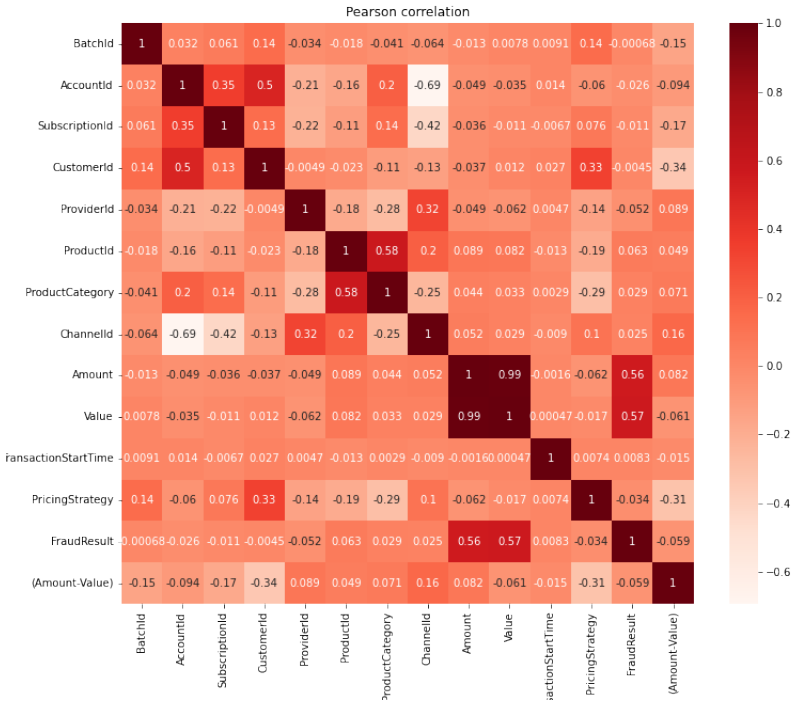


Figure 1: Corrélation de Pearson entre les features

3 Conclusion

Nous avons observé une diminution de la divergence.