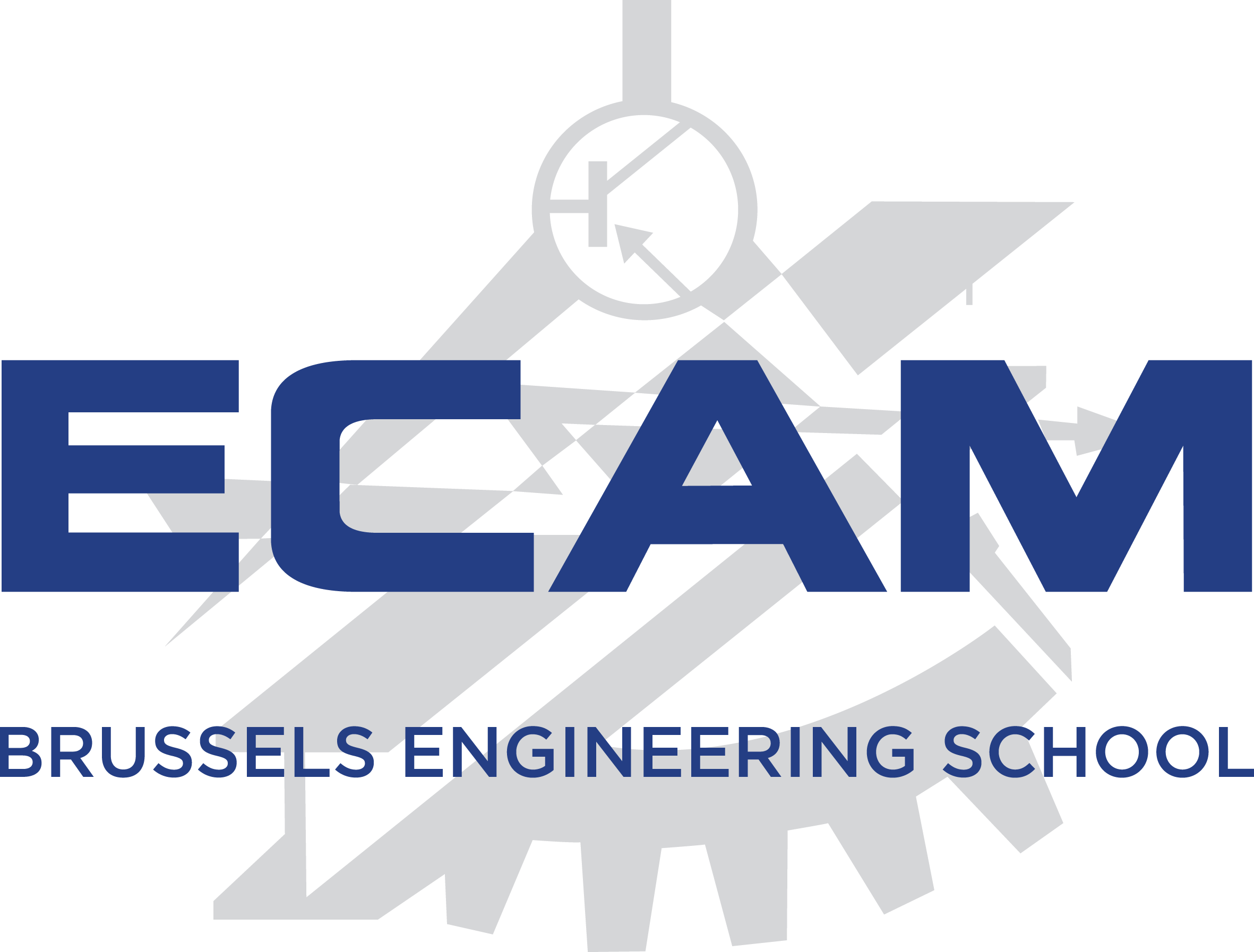
Rapport de laboratoire : Intelligence Artificielle

### Matthias L´eonard et Dawid Krasowski 2022/2023



**Superviseur :** Monsieur HASSELMANN Ken

**Contents**

1. [Introduction](#_bookmark0) 3
2. [Pr´esentation des donn´ees](#_bookmark1) 3
   1. [Description des donn´ees](#_bookmark2) 3
   2. [identification des Features](#_bookmark3) 3
   3. [Optimization de la Feature TransactionStartTime](#_bookmark4) 5
3. [Unbalanced Data](#_bookmark5) 6
4. [Conclusion](#_bookmark6) 6

# Introduction

Dans le cadre des laboratoires d’intelligence artificielle dispensé `a l’ECAM en 2`eme Master en ingénieur informatique. Sous la supervision de Monsieur HAS- SELMANN notre projet se base sur la recherche de fraude à la carte ban- caire. Nous travaillons sur un jeu de données de transactions bancaires issues d’Ouganda ces donn´ees on ´et´e fournis lors d’une compétition Xente de 2019. <https://zindi.africa/competitions/xente-fraud-detection-challenge>

Notre projet est disponible sur GitHub `a l’adresse suivante :

<https://github.com/LeTouristeDeLECAM/Lab_AI_Fraud_Detection>

Pour des questions de propri´et´e et droit les données ne sont pas disponibles sur GitHub.

# Pr´esentation des donn´ees

## Description des donn´ees

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

Table 1: Description des donn´ees

## Identification des Features

Dans un premier temps nous cherchons `a identifier les features qui vont nous permettre de cr´eer un mod`ele de pr´ediction de fraude.

Nous pouvons observer que certaines donn´ees ne sont pas utile pour obtenir un mod`ele. Nous d´ecidons 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 ´egalement imaginer `a premi`ere vue que les donn´ees Amount et Value sont similaires. N´eanmoins suite `a une analyse:

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

nous pouvons observer que les deux colonnes ne sont pas totalement iden- tiques.

|  |  |
| --- | --- |
| 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 diff´erence entre Amount et Value

Nous avons d´ecid´e de garder les donn´ees Amount et Value. Car sur les 193 fraudes que comporte le jeu de donn´ees, 17 fraudes sont r´ealis´e quand Amount et Value sont diff´erents (8,8%).

**Features `a supprimer :** Nous pouvons observer que productCategory et pro- ductID sont fortement corr´el´es. il en est de mˆeme pour amount et value.

Pour pousuivre notre analyse nous r´ealisons une analyse en composante prin- cipale (ACP) sur les donn´ees.

Cette analyse nous permet de r´eduire la dimensionnalit´e de nos donn´ees et iden- tifier les donn´ees qui sont les plus importantes.

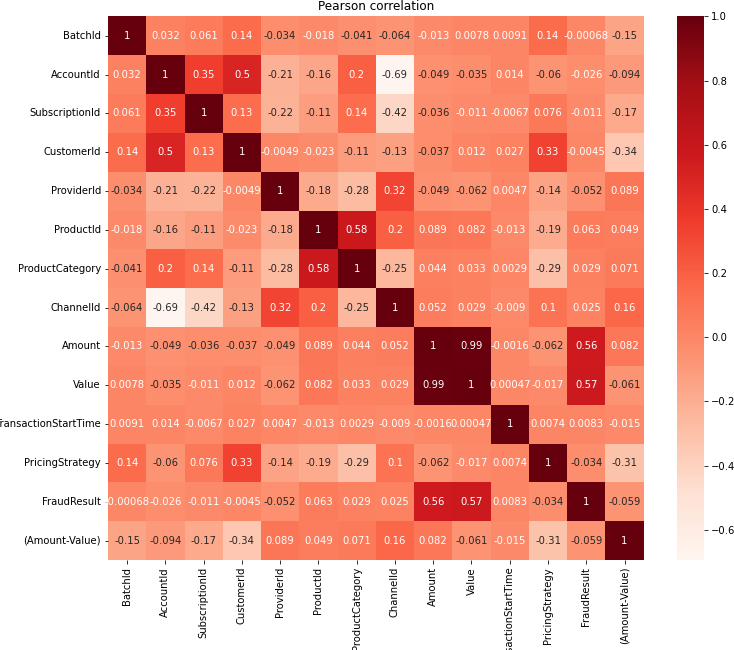


Figure 1: Corr´elation de Pearson entre les features

## Optimization de la Feature TransactionStartTime

La feature TransactionStartTime est un horodatage, les donn´ees disponibles s’´etendent de 15 novembre 2018 au 13 f´evrier 2019, Nous pouvons difficilement d´ecouper cette transformer cette feature pour d´eterminer un comportement sur une p´eriode tel que les mois ou les jours.

Nous avons d´ecid´e d’utiliser uniquement l’heure de la transaction pour identifier un comportement une tendance de comportement.

En observant les graphiques ci-dessus nous pouvons observer que la distri- bution des transactions suit une courbe de gauss avec un creux durant le temps de midi. Nous observons que la distribution des transactions frauduleuses ne suit pas exactement la mˆeme distribution, La distribution semble plus uniforme est semble moins suivre une courbe de gauss avec un plus grand ´ecart-type.

Pour donner plus de sens au observation nous nous int´eressons `a la proportion de transaction frauduleuse en fonction de l’heure.

Nous pouvons observer que la proportion de transaction frauduleuse est plus

´elev´ee durant la nuit et le temps de midi.

Nous avons d´ecid´e de transformer la feature TransactionStartTime en une fea- ture cat´egorielle:

* 1: Morning 04h00 - 11h59
* 2: Lunch 12h00 - 13h59
* 3: Afternoon 12h00 - 19h59
* 4: Night 20h00 - 03h59

Observons la distribution des transactions avec la nouvelle feature.

Avec la transformation de la feature TransactionStartTime en une feature cat´egorielle nous observons que la proportion de transaction frauduleuse est plus

´elev´ee durant la nuit et le temps de midi.

Avec cette transformation nous avons accrue l’importance de cette features pour notre mod`ele.

# Unbalanced Data

Pour palier au probl`eme de donn´ees d´es´equilibr´ees il nous est possible de tra- vailler avec plusieurs m´ethodes. Nous pouvons utiliser des m´ethodes d’Oversampling ou d’Undersampling ou encore une combinaison des deux.

Nous avons d´ecider d’utiliser la m´ethode easy ensemble qui est une m´ethode d’undersampling al´eatoire coupler avec un algorithme de boosting. Conceptuelle- ment cette m´ethode consiste `a cr´eer plusieurs sous-ensembles de donn´ees en sous-

´echantillonnant la classe majoritaire. Nous utilisons Adaboost comme classifieur pour chaque sous-ensemble de donn´ees, Nous utiliseons ´egalement comparons

´egalement avec l’algorithme Gradient Boosting Classifier. Nos r´esultats sont pr´esent´es dans le tableau ci-dessous.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Recall | F1-Score |
| AdaBoost | 0.9995 | 0.8571 | 0.8571 |
| AdaBoost | 0.9995 | 0.8571 | 0.8571 |
| Gradient Boosting | 0.9995 | 0.8571 | 0.8571 |

Table 3: R´esultats de la m´ethode Easy Ensemble Classifier

# Conclusion

Nous avons observ´e une diminution de la divergence.

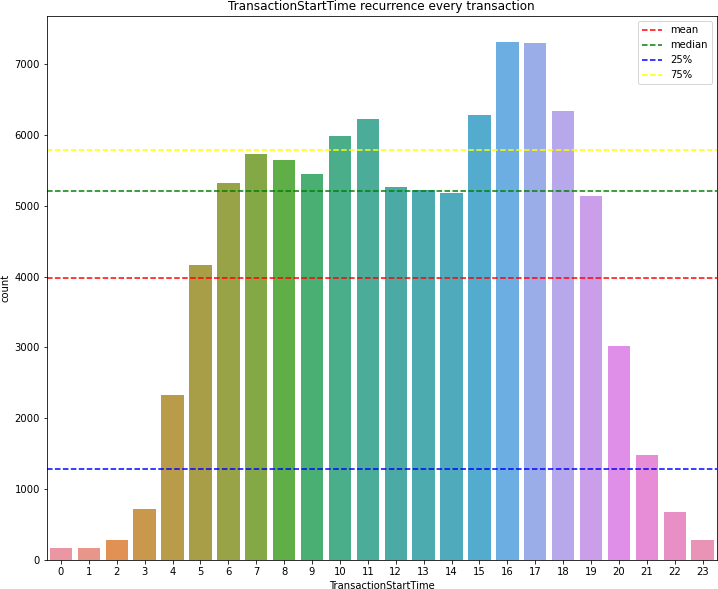


Figure 2: Distribution des transactions en fonction de l’heure

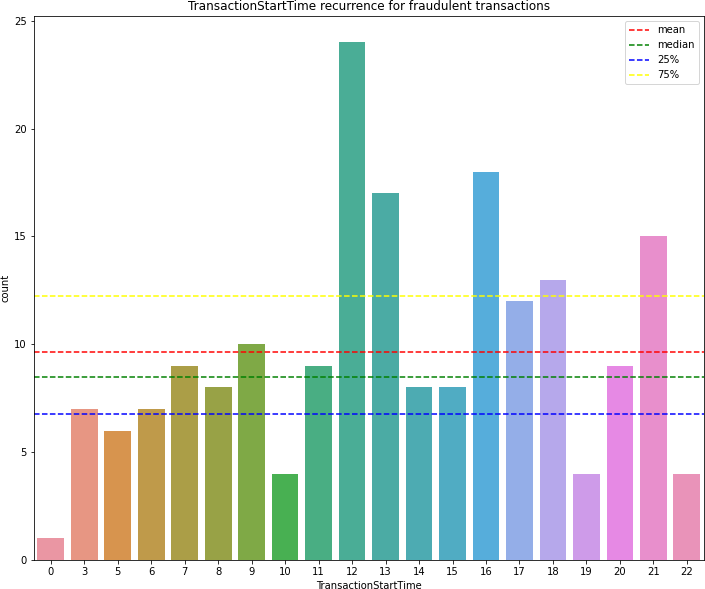


Figure 3: Distribution des transactions frauduleuses en fonction de l’heure

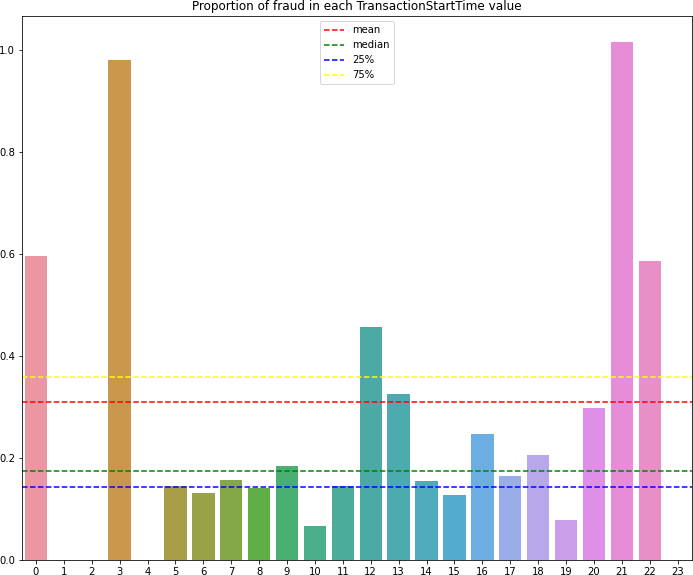


Figure 4: Distribution du rapport des transactions frauduleuses en fonction de l’heure

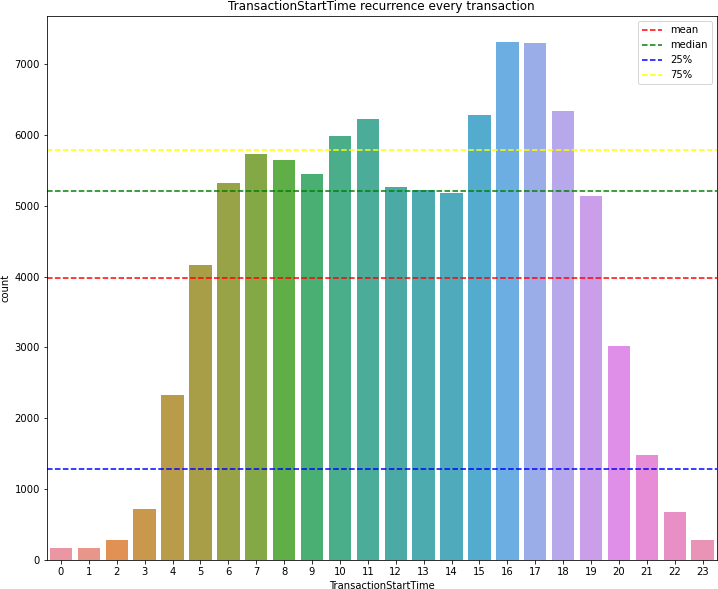
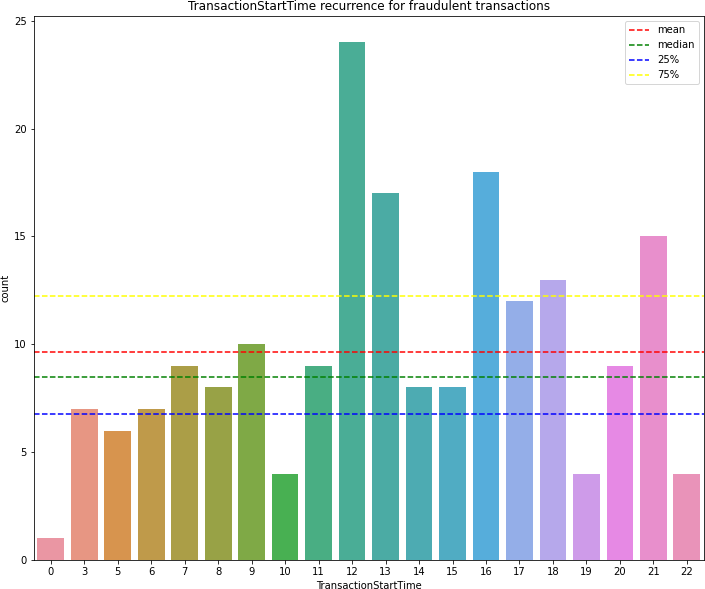
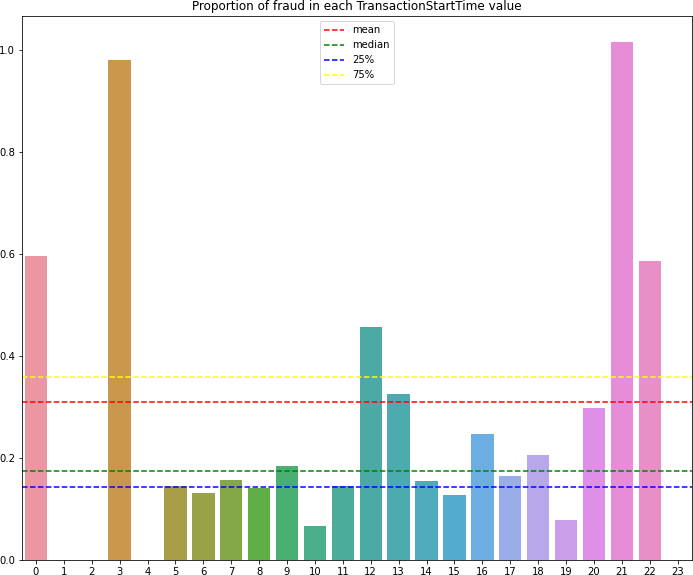
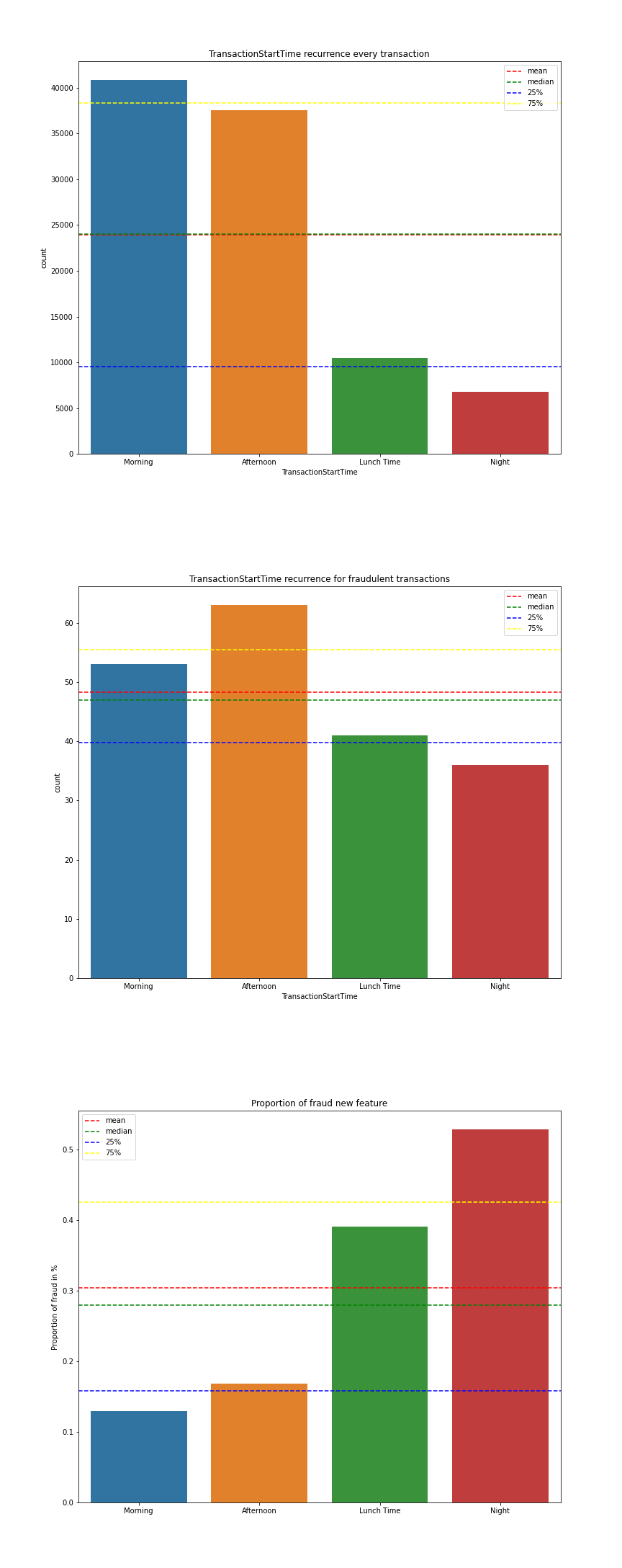
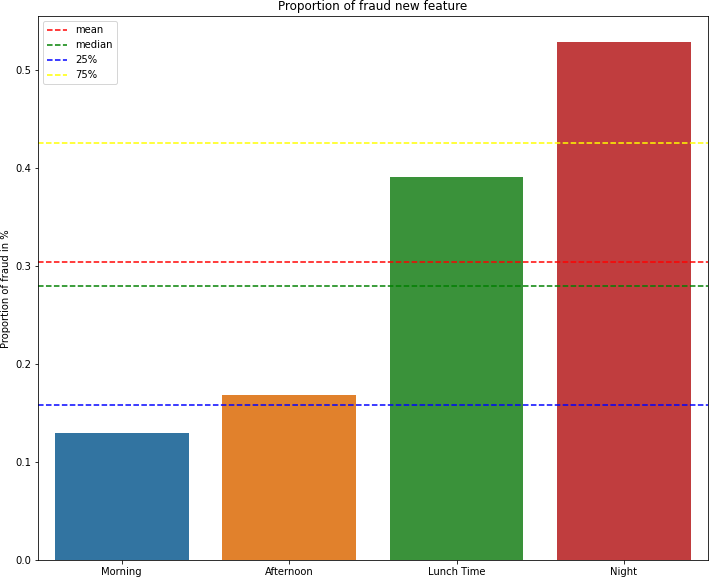
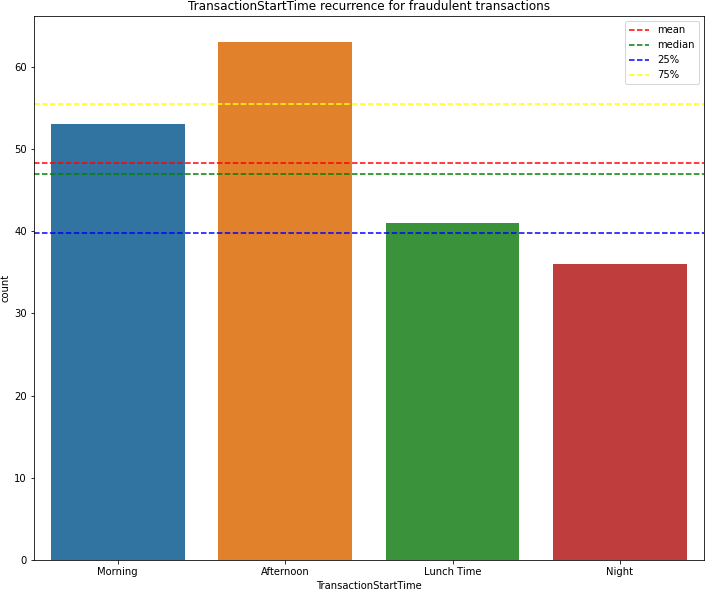
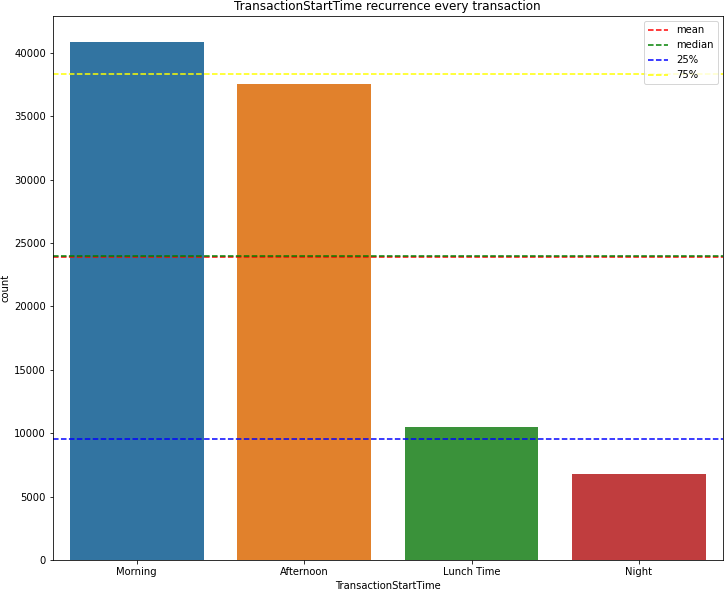
  

Figure 5: Some grouped images





(a)

Figure 7: Distribution des transactions en fonction de l’heure: toutes (a), fraud- uleuses (b), proportion frauduleuses (c)