Mamadou Issa NDIAYE | Machine Learning | November 18, 2023

Détection de Fraude par Carte de Crédit

Master 1 en Data Science & IA



Tables de matières

[Introduction 2](#_Toc151745796)

[Anatomie d’une fraude à la carte bancaire et Répartition des fraudes bancaires : 3](#_Toc151745797)

[Collecte de données : 4](#_Toc151745798)

[Exploration de données : 4](#_Toc151745799)

[Prétraitement de données : 6](#_Toc151745800)

[Modélisation : 6](#_Toc151745801)

[Evaluation du modèle : 7](#_Toc151745802)

[Optimisation 9](#_Toc151745803)

[Interprétation 10](#_Toc151745804)

[Conclusion : 10](#_Toc151745805)

# Introduction

Ce rapport sera divisé en de parties dont elles feront ressortir le travail effectué dessus et bien d’autres aspects permettant de mieux comprendre le projet.

Le Machine Learning étant de donner à un ordinateur ou une machine, la capacité d’apprendre sans programmer de manière explicite.

La détection de fraude par carte de crédit est une composante cruciale dans la sécurité des transactions financières. Avec l'essor des transactions électroniques, la nécessité de détecter et de prévenir la fraude par carte de crédit est devenue une priorité majeure pour les institutions financières et les entreprises.

La régression logistique, une technique de Machine Learning, est largement utilisée pour résoudre ce problème en raison de sa capacité à modéliser et prédire des événements binaires, tels que la fraude ou la non-fraude, en se basant sur des caractéristiques ou des variables explicatives.

Le processus de détection de fraude par carte de crédit en utilisant la régression logistique commence par la collecte de données historiques comportant des informations sur les transactions, telles que le montant de la transaction, la localisation, l'heure de la transaction, le type de commerçant, etc. Ces données servent de base à l'entraînement du modèle.

Ensuite, ces données sont nettoyées et préparées pour l'analyse. Cela peut inclure l'élimination des valeurs aberrantes, la normalisation des données, la gestion des valeurs manquantes, et la création ou l'ingénierie de nouvelles caractéristiques pertinentes pour améliorer la performance du modèle.

Une fois les données préparées, le modèle de régression logistique est entraîné à partir de ces données historiques étiquetées comme frauduleuses ou non. Le modèle cherche à établir une relation entre les caractéristiques des transactions et la probabilité qu'une transaction soit frauduleuse.

Lors de la phase de test, le modèle est évalué sur des données qu'il n'a pas vues auparavant. Il prédit la probabilité de fraude pour chaque nouvelle transaction et applique un seuil pour classer les transactions en frauduleuses ou non. Ce seuil est généralement ajusté pour trouver le bon équilibre entre la sensibilité (capacité à détecter la fraude réelle) et la spécificité (capacité à éviter les fausses alertes).

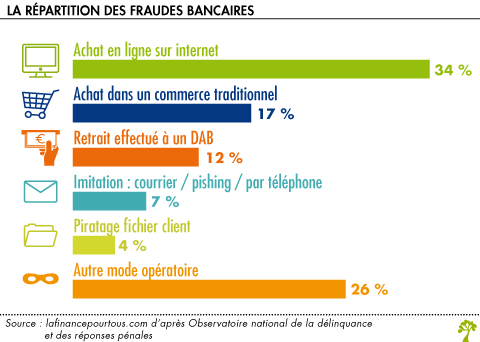
La régression logistique présente des avantages pour la détection de fraude par carte de crédit, notamment sa capacité à gérer des ensembles de données massifs, sa rapidité d'exécution et sa facilité d'interprétation des résultats. Cependant, il est important de noter qu'elle n'est pas la seule méthode utilisée pour cette tâche, et souvent, une approche combinée de plusieurs techniques de Machine Learning est adoptée pour améliorer la précision et la robustesse de la détection de fraude.

# Anatomie d’une fraude à la carte bancaire et Répartition des fraudes bancaires :

Cette image ci-dessous nous montre comment se fait la fraude par carte bancaire pour comprendre de manière explicite et concise.



Ci-dessous, on a une image qui évoque la répartition des fraudes bancaires qui sont connaissables.



# Collecte de données :

Le dataset utilisé pour ce projet est disponible sur Kaggle. Il contient 284 807 transactions, dont 492 sont frauduleuses. Le jeu de données ne contient que des variables d’entrée numériques qui sont le résultat d’une transformation PCA. Malheureusement, en raison de problèmes de conﬁdentialité, nous ne pouvons pas avoir accès aux fonctionnalités d’origine et plus d’informations générales sur les données. Les fonctionnalités V1, V2, . . . V28 sont les principaux composants obtenus avec PCA, les seules fonctionnalités qui n’ont pas été transformées avec PCA sont « Time » et « Amount ». La fonction « Time » contient les secondes écoulées entre chaque transaction et la première transaction de l’ensemble de données. La fonction « Amount » est le montant de la transaction. La caractéristique « Class » est la variable de réponse et prend la valeur 1 en cas de fraude et 0 sinon. L’objet qu’on cherche à prédire est la caractéristique « Class » qui est une variable booléenne, donc c’est un problème de classiﬁcation.

# Exploration de données :

Une analyse préliminaire des données montre que les transactions frauduleuses sont très rares. En effet, elles ne représentent que 0,17 % du total des transactions. Cela signifie que les modèles de détection de fraude doivent être très précis pour être efficaces. L'exploration des données montre également que certaines caractéristiques sont plus susceptibles d'être associées aux transactions frauduleuses. Par exemple, les transactions frauduleuses ont tendance à être de montants plus élevés, à avoir lieu dans des magasins ou des pays exotiques, et à être effectuées à des heures inhabituelles.

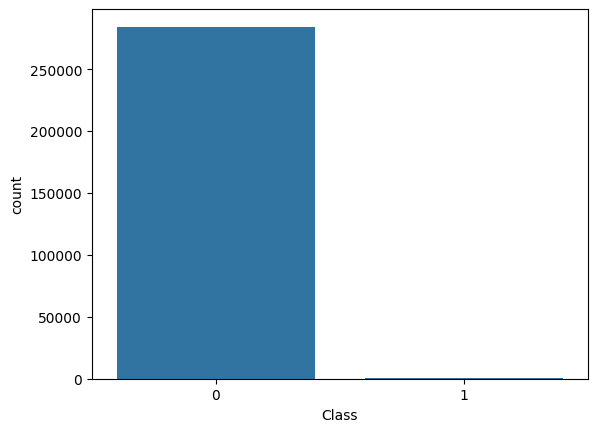


Figure 1 : Visualisation du nombre de répartitions des transactions

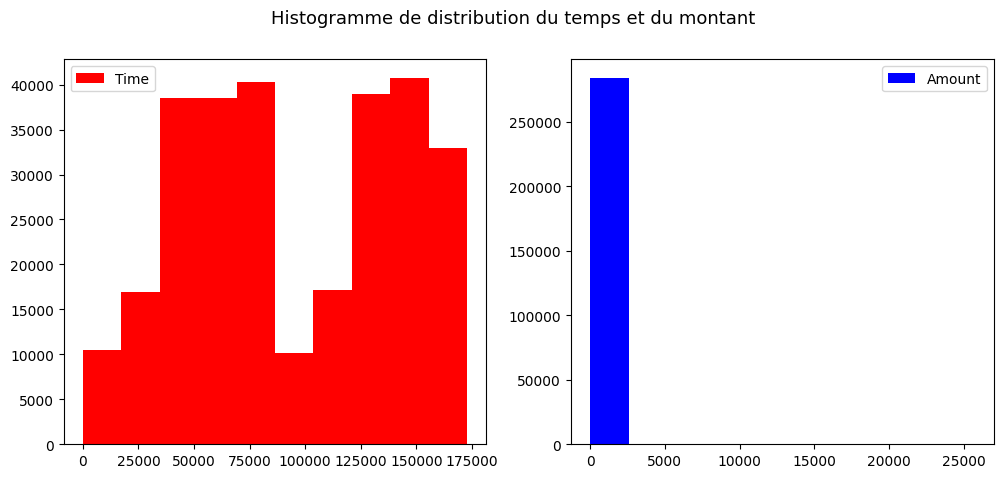


Figure 2 : Histogramme de distribution du temps et du montant pour le dataset principal.

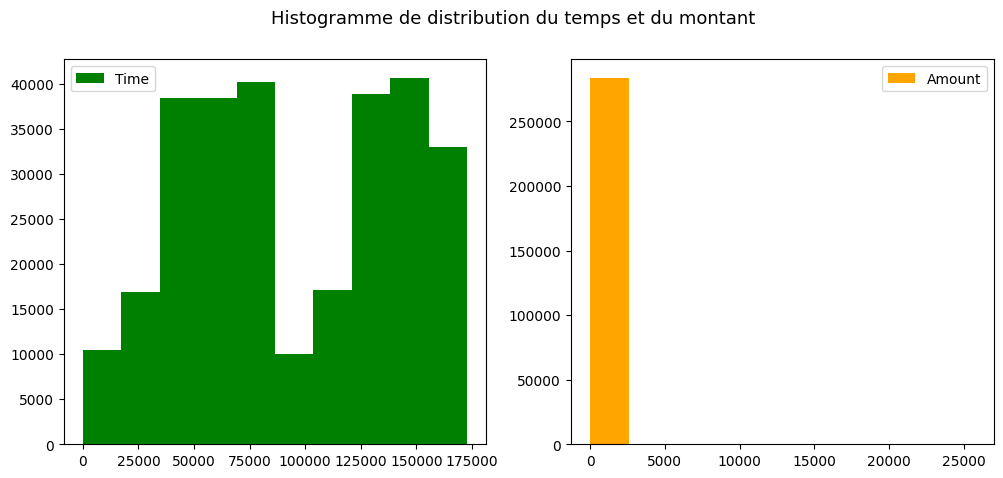


Figure 3 : Histogramme de distribution du temps et du montant pour les transactions normales.

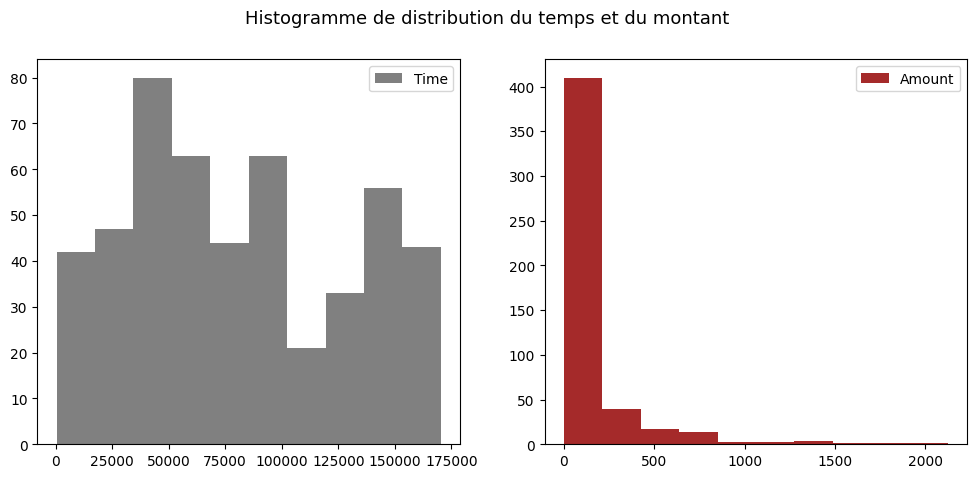


Figure 4 : Histogramme de distribution du temps et du montant pour les transactions frauduleuses.

# Prétraitement de données :

Les données doivent être prétraitées avant d'être utilisées pour entraîner un modèle de Machine Learning. Cela peut inclure des étapes telles que la normalisation, la gestion des valeurs manquantes, et la création de nouvelles caractéristiques. Dans ce cas, la normalisation des données est importante car les différentes caractéristiques ont des échelles différentes. La gestion des valeurs manquantes est également importante, car le dataset pourrait avoir des valeurs manquantes pour certaines caractéristiques mais dans notre cas non. Également nous avons procéder à un échantillonnage pour avoir le même nombre de transactions pour faire notre prédiction. Toutefois, on pouvait utiliser l’ensemble des données sans pour autant l’échantillonner.

# Modélisation :

Plusieurs algorithmes de Machine Learning peuvent être utilisés pour la détection de fraude à la carte de crédit. Les algorithmes les plus courants sont la régression logistique, les arbres de décision, et les réseaux de neurones. Dans ce projet, nous allons utiliser la régression logistique. La régression logistique est un algorithme de classification qui est bien adapté aux problèmes de déséquilibre des classes, comme la détection de fraude.



Figure 5 : Mise en place de notre modèle.

# Evaluation du modèle :

La performance du modèle doit être évaluée avant de l'utiliser en production. Cela peut être fait en utilisant des mesures telles que la précision, le rappel, la F1-score, et la courbe ROC. Dans ce projet, nous allons utiliser la précision, le rappel, la F1-score et la courbe ROC pour évaluer le modèle. La précision mesure le pourcentage de transactions frauduleuses correctement identifiées. Le rappel mesure le pourcentage de transactions frauduleuses qui ont été identifiées par le modèle. La F1-score est une combinaison de la précision et du rappel. La courbe ROC est créé en traçant le taux de vrais positifs (TPR) par rapport au taux de faux positifs (FPR) à divers réglages de seuil de classiﬁeur utilisé.

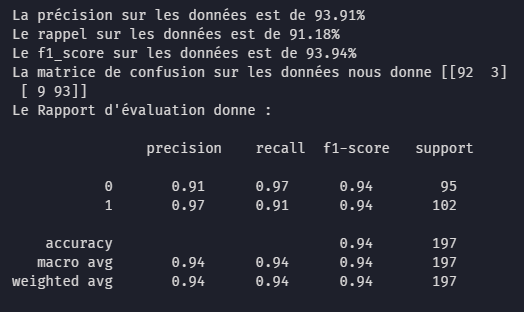


Figure 6 : Evaluation du modèle à travers les métriques.

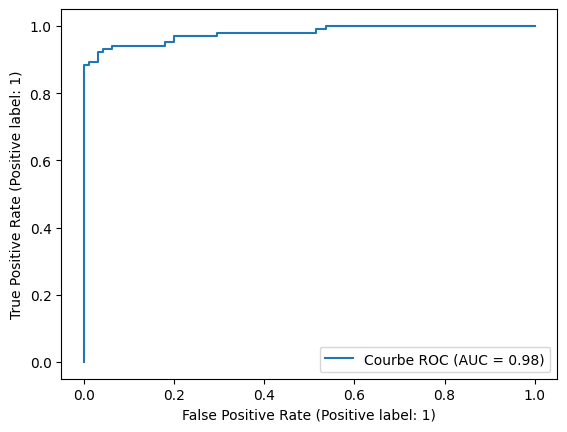


Figure 7 : Courbe ROC.

# Optimisation

Une fois que le modèle a été évalué, il peut être optimisé pour améliorer sa performance. Cela peut être fait en ajustant les hyperparamètres du modèle, en réduisant la dimensionnalité des données, ou en utilisant un ensemble de modèles. Ainsi on procédera à une évaluation du modèle optimisé.

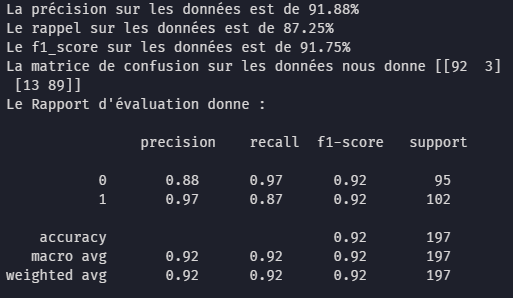


Figure 8 : Evaluation du modèle optimisé.

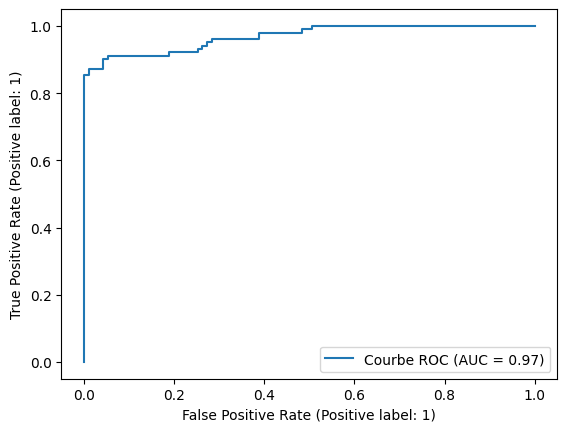


Figure 9 : Courbe ROC du modèle optimisé.

# Interprétation

Il est important d'interpréter les prédictions du modèle pour comprendre comment il fonctionne. Cela peut aider à identifier les caractéristiques importantes qui sont utilisées par le modèle pour prendre des décisions.

# Conclusion :

La détection de fraude à la carte de crédit est un problème complexe. Les modèles de Machine Learning peuvent être utilisés pour détecter les fraudes, mais ils doivent être bien conçus et optimisés pour être efficaces. Ce projet vise à développer un modèle de Machine Learning capable de détecter les fraudes à la carte de crédit. Le projet a suivi les étapes décrites ci-dessus pour collecter, explorer, prétraiter, modéliser, évaluer, optimiser et interpréter les données.

Comme suggestions pour améliorer le notebook :

* Ajout d’une section sur la sélection des caractéristiques ;
* Ajour d’une section sur la validation croisée et
* Ajout d’une section sur l’interprétation de la courbe ROC.

On peut également utiliser d’autres algorithmes de Machine Learning pour la détection de fraude par carte de crédit comme par exemple : Arbres de décisions ou réseaux de neurones.