# DONNEES MASSIVES ET APPRENTISSAGE AUTOMATIQUE

Sujet du projet: Outil pour prédire les cas de fraudes sur cartes bancaires

UNIVERSITE PARIS NANTERRE
Préparé par: DIALLO MAMADOU

## I) Introduction :

Dans le cadre de notre projet portant sur la matière Données Massives et Apprentissage Automatique, nous avons choisi un sujet sur les cas de fraudes sur cartes bancaires. Aujourd'hui, les cas de fraudes sur cartes bancaires que ce soit par achat internet ou bien utilisation de la carte pour retrait ont explosé. Il devient alors une nécessité de fournir aux institutions financières des outils pour détecter les tentatives et de prévenir les fraudes sur cartes bancaires. Dans cette logique, nous étions allés chercher une base de données sur Kaggle portant sur les cas de fraudes sur cartes bancaires.

## II) Méthodologie :

La base de données de départ était constituée d'un million d'observations que nous avons rétrécit jusqu'à mille observations pour des soucis de capacité de calcul liés au données massives. Le travail de traitement de données n'a pas été pénible puisque nous sommes tombés sur une base de données prés à être utilisés ou il n'y avait pas de données manquantes. Cependant, nous avons également pris la liberté de faire une analyse descriptive de la base de données afin de savoir la répartition statistique de variables d'études, à voir en annexe.

L'étude a été mené grâce à des méthodes de machine Learning comme : la cross validation, le forêt aléatoire et l'arbre de décision.

## III) Résultats :

En termes de résultats, nous avons constaté que les trois algorithmes ont données des scores de fraudes sur cartes bancaires différents mais également des résultats différents avec la matrice de confusion.

La matrice de confusion de l'arbre de décision fournit les résultats suivants :

Sur la matrice(1,1), sur 300 utilisations de carte bancaire, l'algorithme prédit 279 non-fraudes. La matrice(2,1), l'algorithme prédit une fraude alors qu'il y'a pas de fraude. La matrice(1,2), l'algorithme prédit non-fraude alors qu'il y'a une fraude. La matrice(2,2), l'algorithme prédit 19 fraudes sur 300 utilisations.

La matrice de confusion du random forest fournit les résultats suivants :

Sur la matrice(1,1), sur 300 utilisations de carte bancaire, l'algorithme prédit 277 non-fraudes. La matrice(2,1), l'algorithme prédit une fraude alors qu'il y'a pas de fraude. La matrice(1,2), l'algorithme prédit non-fraude alors qu'il y'a 3 une fraude. La matrice(2,2), l'algorithme prédit 19 fraudes sur 300 utilisations.

A propos du score, l'arbre de décision fournit le meilleur score prés de 99,333%, puis le random forest avec 98,677% et enfin le cross validation 97 ,75%.

IV) Conclusion :

La machine Learning est devenue aujourd'hui un outil essentiel dans la prise de décision des acteurs de la finance et de la banque car elle permet de faire des prédictions à partir de données financières ou bancaires gigantesques. En ce sens l'apprentissage automatique permet des prises de décisions rapide et efficace afin d'endiguer les transactions suspects ou frauduleuses, ce qui peut aider à réduire les pertes liées à la fraude. Ce pendant fournir l'algorithme qui permet de minimiser les faux positifs et les faux négatifs demeure sujet principal de recherche de l'apprentissage automatique. Le choix sur les algorithmes a utilisé dépend des objectifs du modélisateur et de la justesse dans la prédiction en réduisant le plus possible de biais. Dans le cas de notre projet, nous avons choisi de challenger trois algorithmes et à la fin nous avons constaté que l'arbre de décision fournit le meilleur score.

ANNEXES :

**TABLEAU DES VARIABES D'INTERETS** :

| | distance_from_home | distance_from_last_transaction | \ |
|---|---|---|---|
| 0 | 57.877857 | 0.311140 | |
| 1 | 10.829943 | 0.175592 | |
| 2 | 5.091079 | 0.805153 | |
| 3 | 2.247564 | 5.600044 | |
| 4 | 44.190936 | 0.566486 | |
| .. | ... | ... | |
| 995 | 9.873417 | 1.022586 | |
| 996 | 168.091704 | 6.304360 | |
| 997 | 44.047622 | 0.510298 | |
| 998 | 2.998418 | 0.193681 | |
| 999 | 38.133449 | 0.167059 | |

| | ratio_to_median_purchase_price | repeat_retailer | used_chip | \ |
|---|---|---|---|---|
| 0 | 1.945940 | 1.0 | 1.0 | |
| 1 | 1.294219 | 1.0 | 0.0 | |
| 2 | 0.427715 | 1.0 | 0.0 | |
| 3 | 0.362663 | 1.0 | 1.0 | |
| 4 | 2.222767 | 1.0 | 1.0 | |
| .. | ... | ... | ... | |
| 995 | 0.346643 | 1.0 | 0.0 | |
| 996 | 0.416833 | 1.0 | 1.0 | |
| 997 | 0.624706 | 1.0 | 0.0 | |
| 998 | 0.743416 | 1.0 | 1.0 | |
| 999 | 3.965637 | 1.0 | 0.0 | |

| | used_pin_number | online_order | fraud |
|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 |
| 1 | 0.0 | 0.0 | 0.0 |
| 2 | 0.0 | 1.0 | 0.0 |
| 3 | 0.0 | 1.0 | 0.0 |
| 4 | 0.0 | 1.0 | 0.0 |
| .. | ... | ... | ... |
| 995 | 0.0 | 0.0 | 0.0 |
| 996 | 0.0 | 0.0 | 0.0 |
| 997 | 0.0 | 1.0 | 0.0 |

| | | | |
|---|---|---|---|
| **998** | **0.0** | **1.0** | **0.0** |
| **999** | **0.0** | **1.0** | **0.0** |

[1000 rows x 8 columns]

**Statistiques descriptives:**

| | distance_from_home | distance_from_last_transaction | ratio_to_median_purchase_price | repeat_retailer | used_chip | used_pin_number | online_order | fraud | |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 1000.000000 | 1000.000000 | 1000.000000 | 1000.000000 | 1000.000000 | 1000.000000 | 1000.000000 | 1000.000000 | |
| **mean** | 26.887864 | 5.418143 | 1.656683 | 0.893000 | 0.314000 | 0.117000 | 0.673000 | 0.079000 | |
| **std** | 63.868699 | 35.597907 | 2.349131 | 0.309268 | 0.464349 | 0.321581 | 0.469352 | 0.269874 | |
| **min** | 0.104184 | 0.001448 | 0.016933 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | |
| **25%** | 4.006805 | 0.306441 | 0.468934 | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | |
| **50%** | 10.346908 | 0.955936 | 0.958947 | 1.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | |
| **75%** | 25.467761 | 3.112754 | 1.922379 | 1.000000 | 1.000000 | 0.000000 | 1.000000 | 0.000000 | |
| **max** | 965.910612 | 990.070315 | 36.074366 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | |