Mathis TRUONG et Timothé WINKLER

RAPPORT

1 - Importation des données et exploration initiale (EDA)

Le jeu de données analysé provient d'une simulation de campagne de phishing dans laquelle plusieurs utilisateurs se sont vu recommander un produit via un canal spécifique (mail, Instagram ou Facebook).

Chaque entrée contient :des scores d'intérêt (gaming, design Instagram, football),

- l'âge de l'utilisateur,
- un produit recommandé (FIFA, Fortnite, Instagram Pack),
- un canal de diffusion,
- le succès ou non de la campagne.

L'importation a été effectuée avec pandas.read_csv() en utilisant le séparateur **;** . Les premières lignes et le résumé statistique ont confirmé une structure cohérente et l'absence de colonnes fusionnées.

Nous avons optimisé l'importation en utilisant le bon séparateur, en convertissant les colonnes dans des types plus légers et en nettoyant rapidement les données incohérentes (*float32*, *category*, *int8/int32*), ce qui a permis de réduire l'utilisation mémoire et d'accélérer l'ensemble du traitement.

Notre première analyse révèle :

- des scores généralement compris entre 0 et 100,
- quelques valeurs aberrantes dépassant largement ces bornes,
- des incohérences mineures dans les textes ("Fornite" au lieu de "Fortnite", variations dans les canaux),
- quelques valeurs manquantes sur le score football, l'âge ou le produit.

2 - Nettoyage et Mise en Forme des Données

Après l'importation du fichier, nous avons réalisé un ensemble d'opérations de nettoyage afin de garantir la cohérence, la qualité et l'exploitabilité du dataset. Cette étape était essentielle avant toute analyse statistique ou visualisation.

Plusieurs colonnes étaient initialement importées sous le type object, ce qui n'est pas optimal pour réaliser des calculs ou des groupements.

Nous avons donc procédé aux ajustements suivants :

- Conversion de campaign_success en valeurs numériques 0/1 afin de faciliter le calcul du taux de réussite.
- Uniformisation des colonnes textuelles (recommended_product, canal_recommande) en minuscules et en supprimant les espaces superflus.
- Correction d'incohérences typographiques, par exemple "Fornite" remplacé par "Fortnite".
- Transformation des scores (gaming_interest_score, insta_design_interest_score, football_interest_score) en float32 et de l'âge en int8.

Ces optimisations permettent un gain de mémoire tout en garantissant un format compatible avec les analyses.

Avant nettoyage, plusieurs entrées présentaient des valeurs manquantes, notamment sur :

- le score football,
- l'âge,
- le produit recommandé.

Comme ces informations sont indispensables pour les analyses, nous avons choisi de supprimer les lignes incomplètes.

Afin d'assurer la cohérence des scores, nous avons repéré les valeurs en dehors de l'intervalle logique 0,100.

Les anomalies détectées incluait des points extrêmes, notamment des scores :

- négatifs,
- supérieurs à 400, 500, voire 700 selon les colonnes.

Ces valeurs sont visibles dans les graphiques générés :

- anomalies_gaming_interest_score.png
- anomalies_insta_design_interest_score.png
- anomalies_football_interest_score.png

Nous avons choisi de supprimer toutes les lignes contenant au moins une anomalie afin d'obtenir un dataset propre et réaliste.

Pour faciliter les analyses statistiques, nous avons enrichi le dataset en ajoutant plusieurs segmentations :

- **age_group** : regroupement des âges en classes pertinentes (18–24, 25–34, 35–44, etc.)
- **gaming_segment** : segmentation des scores gaming en catégories *faible*, *moyen*, *fort*.

Ces nouvelles colonnes ont été converties au type category, ce qui optimise les performances lors des groupements et calculs de KPI.

À l'issue du nettoyage, le dataset est complet, cohérent et prêt pour l'analyse : les valeurs manquantes critiques ont été supprimées, les données numériques ont été corrigées et uniformisées, les types ont été optimisés et les segments d'analyse clairement définis. L'ensemble forme une base solide pour calculer les KPI, produire des visualisations fiables et construire le data telling.

3 - Visualisation et Distribution des Variables

Nous avons représenté la distribution des trois scores d'intérêt (gaming, design Instagram, football) à l'aide d'un histogramme regroupé (distrib_scores.png).

Les courbes montrent que :

- les trois scores sont répartis de manière relativement homogène,
- aucune concentration extrême n'apparaît après le nettoyage,
- les valeurs se situent bien entre 0 et 100, confirmant la cohérence générale des données.

Cette visualisation confirme que le dataset, une fois nettoyé, présente des distributions équilibrées et adaptées à une analyse statistique fiable.

La matrice de corrélation (corr_matrix.png) a été générée afin d'identifier les relations entre les variables quantitatives : scores d'intérêt, âge et réussite de la campagne.

Les résultats mettent en évidence :

- de faibles corrélations entre les scores d'intérêt et la variable campaign_success, ce qui suggère que la réussite d'une attaque ne dépend pas directement d'un seul centre d'intérêt.
- une corrélation modérée entre l'âge et l'intérêt pour le football, indiquant un léger effet générationnel.
- peu de corrélations entre les scores eux-mêmes, les centres d'intérêt semblent indépendants les uns des autres.

La matrice montre donc que les relations entre variables quantitatives sont limitées, ce qui confirme la nécessité d'analyser également les variables catégorielles (produit, canal, tranche d'âge) dans les parties suivantes.

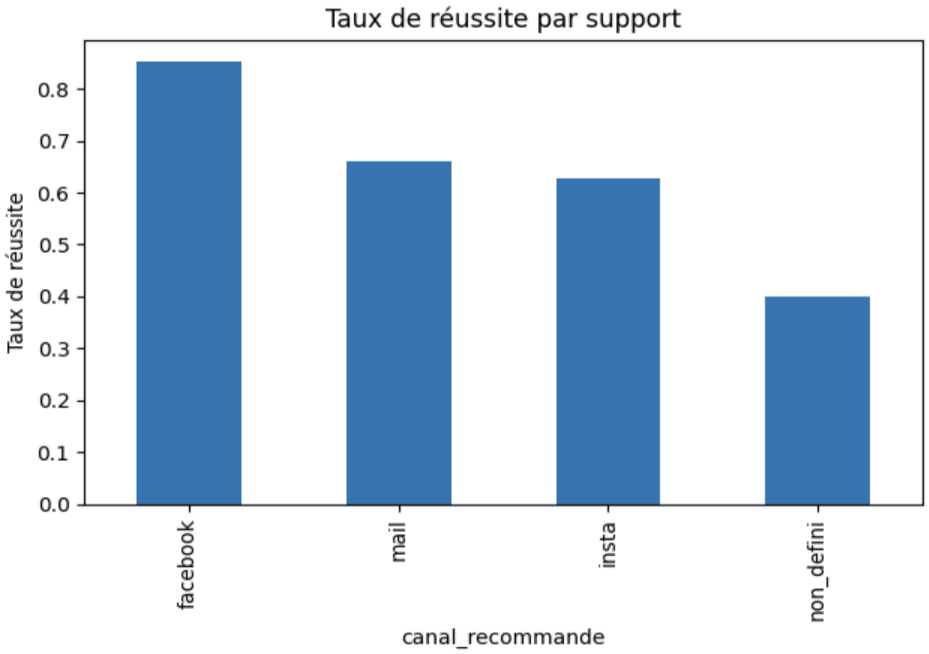
4 - Analyse Statistique et KPI

L'objectif de cette section était d'évaluer l'efficacité de la campagne selon plusieurs dimensions : le produit recommandé, le canal de diffusion, la tranche d'âge des utilisateurs et leur niveau d'intérêt. Les KPI ont été calculés à partir de la variable campaign_success_bool, où 1 indique une réaction positive à la campagne.

Taux de réussite par produit recommandé :

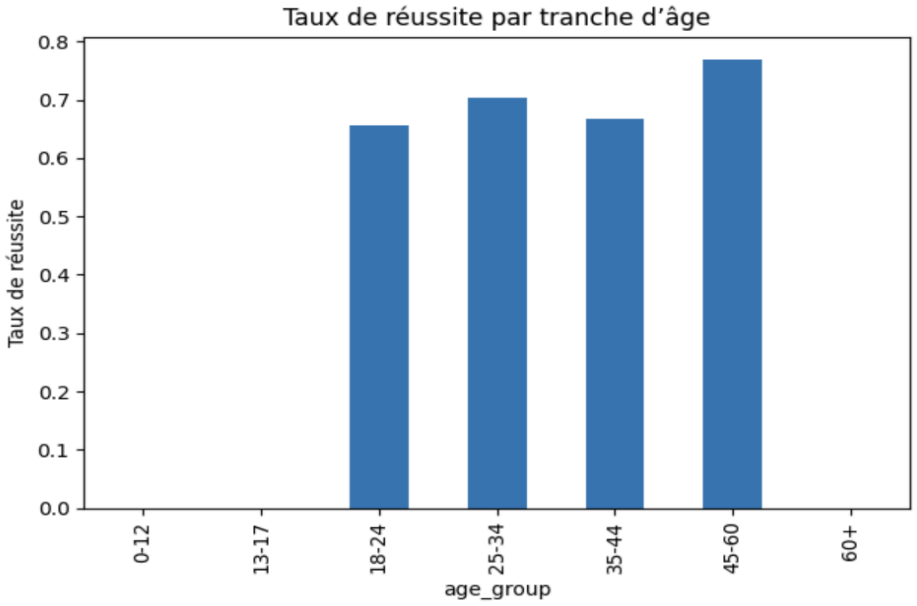
L'analyse montre que certains produits génèrent beaucoup plus d'engagement. FIFA est le produit le plus performant (~72 %), ce qui suggère un fort attrait du public pour ce type de contenu. Cela indique que la thématique choisie pour la recommandation influence directement le succès de la campagne.

Taux de réussite par canal de diffusion :



Facebook est de loin le canal le plus efficace (~85 %), avec un taux de réussite très supérieur aux autres. Cela peut s'expliquer par un public plus réceptif sur ce réseau ou une meilleure visibilité des messages.

Taux de réussite par tranche d'âge :



Le groupe 45–60 ans est le plus vulnérable avec un taux proche de 78 %.
Les tranches jeunes et adultes intermédiaires ont des résultats similaires autour de 66–70 %

Les KPI montrent que :

- Le produit le plus efficace est FIFA, suivi de Fortnite.
- Le canal Facebook surpasse largement les autres supports.
- Les 45–60 ans réagissent beaucoup plus aux campagnes.
- L'intérêt gaming élevé semble également être un facteur de sensibilité.

La réussite d'une campagne de phishing dépend donc fortement de variables qualitatives (produit, canal, tranche d'âge) et non uniquement des scores numériques bruts. Cette analyse fournit des indications précieuses pour identifier les groupes les plus exposés.

5 - Datatteling

En analysant les résultats, on remarque assez vite que certains groupes d'utilisateurs réagissent beaucoup plus que d'autres à la campagne de phishing. Par exemple, la tranche des 45–60 ans affiche un taux de réussite très élevé, autour de 78 %. C'est clairement le groupe le plus sensible de tout le dataset. Les autres tranches tournent plutôt entre 66 % et 70 %, ce qui reste élevé, mais montre qu'il y a vraiment une différence de comportement selon l'âge.
Du côté des supports, Facebook se démarque complètement avec un taux de réussite de 85 %. Les campagnes envoyées par mail ou via Instagram fonctionnent moins bien (autour de 63–66 %). Cela laisse penser que les utilisateurs, surtout les plus âgés, sont plus confiants ou moins méfiants quand un message apparaît directement dans leur fil Facebook.
Enfin, les produits recommandés jouent un rôle important. Les offres liées au gaming, comme FIFA (~72 %) ou Fortnite (~70 %), déclenchent plus de réactions que des contenus plus "classiques" comme un pack Instagram (~66 %). On voit donc que le thème du message influence fortement la façon dont les utilisateurs réagissent.
Dans l'ensemble, les chiffres montrent que la vulnérabilité vient surtout d'une combinaison de facteurs : l'âge, le canal et le type de contenu proposé.

Scénario :

À partir de ces observations, on peut imaginer un scénario d'attaque qui s'appuie directement sur les résultats obtenus. Si un cybercriminel voulait maximiser ses chances, il ciblerait probablement les utilisateurs de 45 à 60 ans, puisqu'on a vu que c'est le groupe qui réagit le plus.
Ensuite, il choisirait naturellement de passer par Facebook, qui est le canal le plus efficace selon nos chiffres. Pour attirer encore plus l'attention, il pourrait créer une fausse publicité autour d'une offre gaming très attractive, par exemple un FIFA Ultimate Pack gratuit ou un bonus spécial Fortnite.
Le message serait présenté comme une opportunité à saisir rapidement, avec un lien à cliquer pour récupérer la récompense. Sur Facebook, ce genre d'annonce peut paraître crédible, surtout si elle ressemble aux publicités habituelles de la plateforme.
Le scénario repose donc sur les trois éléments les plus influents : la cible, le canal et le produit.

Ce scénario est intéressant car il correspond parfaitement à ce que montrent nos chiffres. On sait que les 45–60 ans sont ceux qui réagissent le plus aux campagnes (78 %), ce qui en fait une cible idéale pour une attaque malveillante. En les visant directement, l'attaquant augmente fortement ses chances de succès.
Facebook, avec un taux de réussite d'environ 85 %, est aussi le canal parfait pour diffuser l'attaque. Les utilisateurs ont tendance à accorder plus de confiance aux contenus qu'ils voient sur leur réseau social, ce qui peut réduire leur vigilance.
Enfin, les produits liés au gaming se révèlent très efficaces pour attirer les clics. Entre FIFA et Fortnite, les taux tournent autour de 70 %, ce qui montre que même des personnes de 45–60 ans peuvent être sensibles à ce type de contenu, surtout si celui-ci est présenté comme une offre exclusive ou limitée.
En combinant ces trois éléments, on obtient un scénario de phishing très crédible et potentiellement dangereux. L'intérêt de cette analyse est justement de montrer comment un cybercriminel pourrait s'appuyer sur les données pour cibler précisément les utilisateurs les plus vulnérables.