Business Case

Nom du Projet: Questionnaire visuel pour prévenir de l'usurpation d'identité bancaire

Étudiants: BOUTHEMY Marin & COSTA Jaime & LOUTFI Mehdi
3 étudiants en double-diplôme à l'ENSAE ParisTech (majeure Data Science)

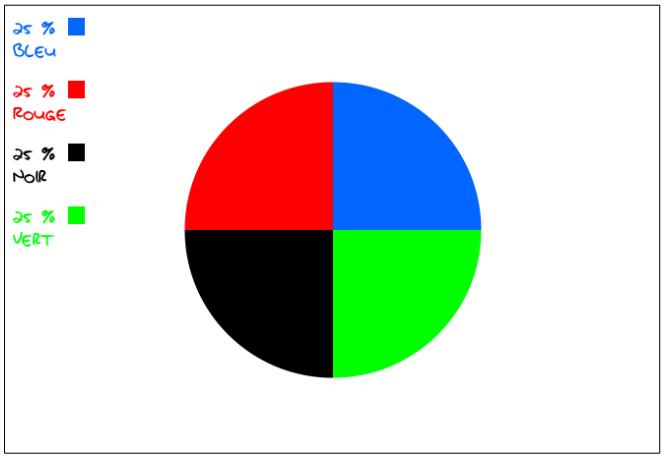
Contexte

De nos jours, lorsqu'on crée un compte bancaire en ligne, il faut au préalable remplir de **nombreuses questions et réponses secrètes** afin de pouvoir confirmer son profil dans le cas d'une tentative d'usurpation d'identité ou d'oubli de mot de passe.
En partant du constat que ces questions **peuvent être difficiles à se souvenir /** peu précises (comme par exemple, "Quel est votre plat préféré ?", "Quel est le nom de votre premier animal?") et ce d'autant plus pour les personnes âgées, notre projet serait de proposer un **système de vérification visuel** et plus aisé à mettre en place.

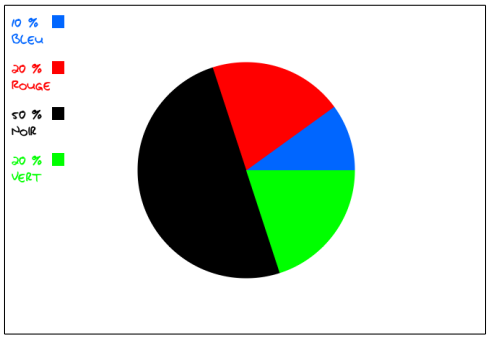
Nous visons ainsi tous les clients utilisant les banques en ligne et qui trouvent fastidieux ou qui ont tendance à **oublier** leurs diverses questions secrètes.

Présentation du Prototype / Utilisation du Machine Learning

Notre projet s'articule ainsi sous la forme de plusieurs questions basées sur le visuel, par exemple à la question de validation quelle est votre couleur préférée, l'utilisateur devra cliquer sur le cercle suivant :



Là où la Data Science entre en jeu est qu'on pourra au préalable effectuer une **étude statistique** afin de déterminer quel est le pourcentage de personnes aimant telle couleur. Ainsi, si il s'avère que seulement 3 % des gens aiment la couleur noire (donc peu de personnes ont cette couleur préférée) on pourra augmenter la taille du cercle de manière à ce qu'une **personne cliquant au hasard** (par exemple un bot ou un hacker) ait une plus grande probabilité de choisir la mauvaise couleur.



A l'inverse, si la couleur bleue est la couleur préférée de la personne, sa taille est réduite de façon à ce qu'on soit sûr que la personne est bien choisie la **couleur qui lui correspond**.

Ainsi, en proposant d'autres questions là aussi basées sur le visuel, par exemple choisir votre film préféré parmi une catégorie de films (horreur / dessins animés / films années 80…) on pourra ainsi s'assurer que les choix restent **cohérents avec le profil de l'utilisateur**.

On peut ainsi supposer qu'une femme aux alentours de 60 ans aura plus tendance à choisir un film ancien plutôt que le dernier Disney comme film préféré. Ainsi, si nous savons que l'utilisateur est une personne âgée, la question aura l'apparence suivante.

Quel est votre film préféré ?:

DISNEY	HORREUR	DISNEY
HORREUR	DISNEY	GRANDE VADROUILLE

Ici, nous avons volontairement augmenté la **fréquence d'apparition des films « non ciblés »** de manière à ce qu'une sélection aléatoire (par un bot ou un hacker ne connaissant pas le profil de la personne) ait plus de chance d'aboutir au mauvais choix de film.

Deep Learning

Pour savoir si à l'issue de toutes ces réponses, le profil de l'utilisateur semble correct ou non, nous pensons utiliser le **machine learning et plus précisément un réseau de neurones** (aussi appelé deep learning).

En entrée (les inputs) nous lui passerons diverses informations telles que :
-profil de la personne (âge, les premiers choix entrées lors de son inscription en ligne)
-les choix demandés lors de la vérification personne

En sortie le réseau nous donnera ainsi la probabilité que les choix de réponses correspondent au **profil de la personne**. Par exemple, si on obtient une ressemblance à plus de 90 %, on peut considérer que la personne est bien la même, tandis que si on obtient moins de 90 % on pourra inviter la personne à justifier son identité d'une autre façon avant de procéder à un virement important ou autre.

Avantage du Deep Learning

N'ayant qu'une idée vague du profil type, nous ne pouvons pas établir un pourcentage de ressemblance très précis, a contrario le deep learning agissant comme une boîte noire est bien plus efficace pour trouver des **similitudes ou différences dans toutes les informations** qu'on lui fournirait.

Son avantage, est qu'une fois correctement entraînée (à l'aide notamment de plusieurs simulations de profils), le deep learning est très **rapide pour la phase de vérification** des profils et peut donc agir immédiatement.
En termes d'avantage, notre solution pourrait ainsi économiser du temps aux clients et de l'argent au service clientèle et d'aide aux particuliers puisque le système de détection des faux profils serait plus performant. Par ailleurs, le cas où les particuliers auraient oublié leur mot de passe et réponse secrète n'existerait plus, et cela sera ainsi une économie de temps et d'argent pour le service d'assistance au client.

Comparaison à l'Offre Existante

Notre solution serait ainsi bien plus ergonomique et intuitive que l'utilisation de questions / réponses dont les utilisateurs ont la **fâcheuse manie de les oublier** voire de ne pas orthographier correctement la réponse qu'il avait préalablement renseignée.

Par ailleurs, dans le cas où le système d'identification se relèverait moins fiable, il pourrait agir de manière très rapide en complément d'un système plus fiable comme la vérification par téléphone ou autre.

Enfin, en découvrant un système innovant et simplifié en ligne, il est possible que des utilisateurs adhèrent davantage à **l'utilisation et la consultation d'un compte en ligne**.

Spécifications Techniques

Le projet et son implémentation serait ainsi à diviser en deux parties :

-la première dédiée à la création et l'entraînement du **neural network** au moyen d'une simulation de profils types et d'une préalable analyse statistique (sur la couleur préférée des français selon l'âge, etc.)
Pour cela, nous pensons coder au moyen de Python et l'utilisation d'API dédiées au Deep Learning comme **Tensorflow ou son module Keras**. De même, l'utilisation de Scikit-Learn peut également être envisagé dans la partie étude statistique des données (voire Spark si il s'agit de données massives)

-la seconde partie concernerait la mise en ligne et l'expérience utilisateur. Pour cela, il faudrait utiliser notamment du JavaScript, pour tout ce qui tourne autour du site web et son interface.

L'avantage d'utiliser le deep learning est que bon nombre de **modèles de réseaux agissant comme des classifiers / détection de profils existent** déjà et qu'il est très aisé de d'adapter leur structure à notre modèle. Ainsi, coder un tel système de détection ne pose pas un problème majeur et peut très bien être mené à bout en 4 mois (du moins un premier réseau assez sommaire qu'il faudra par la suite renforcer).