LUCKY CART - Modélisation d'un prédicteur

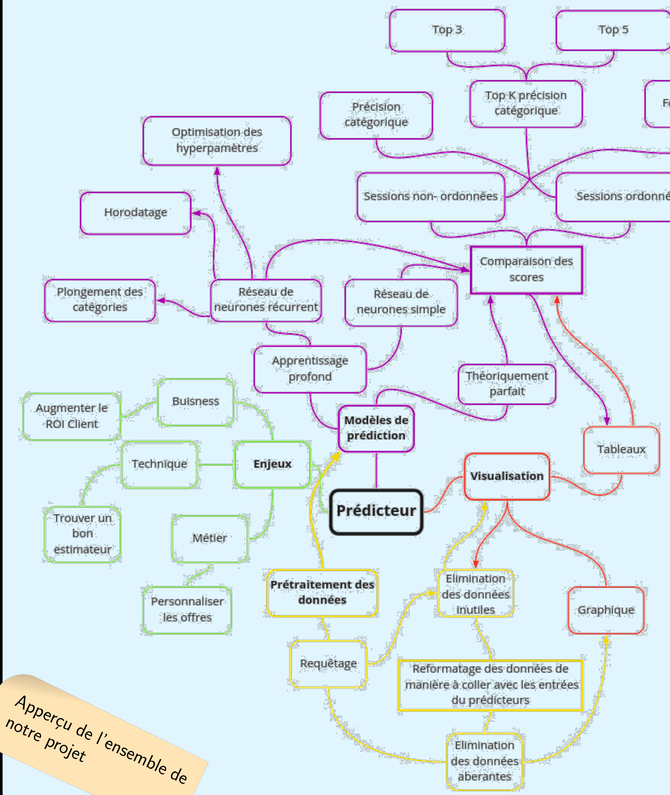
Kenza El M'hamdi - Florian Guily - Naël Sennoun - Aminata Traore - Ruxue Zeng

Polytech Sorbonne
Projet industriel

Présentation de l'entreprise

Lucky Cart est une start-up qui a développé une plateforme basée sur l'intelligence artificielle et a révolutionné le monde de la publicité en ligne en personnalisant les promotions offertes à leurs clients et en utilisant le jeu comme une alternative aux remises classiques : c'est ce que l'on appelle le "Promo Gaming". En tant que futur data scientist, notre enjeu principal est de permettre à la start-up de bien structurer ses données et de les analyser avec les méthodes qui se révèlent les plus efficaces et les plus rapides.
Objectif : Utiliser des méthodes d'apprentissage profond afin de créer un prédicteur. Celui-ci devra, à partir d'un ensemble de données d'entrée d'un utilisateur, ressortir un vecteur de probabilités de ses prochaines actions.

Mind map



Apperçu de l'ensemble de notre projet

Enjeux & impact du projet

Grâce à l'historique des clients, nous arriverons à mieux comprendre leurs besoins, à personnaliser leurs promotions. Le modèle permettra donc de rendre les offres plus impactantes. Ainsi, on pourra augmenter le ROI des entreprises.

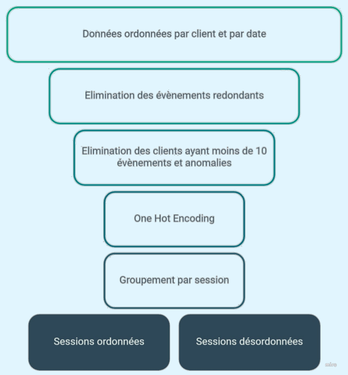
Impacts environementaux

- La surexploitation des ressources naturelles
- La surproduction des déchets.
- La pollution de l'eau et de l'air

Impact sociétaux

- La dépendance à la consommation.
- L'endettement des ménages.
- Le stress et l'anxiété liés aux dettes.

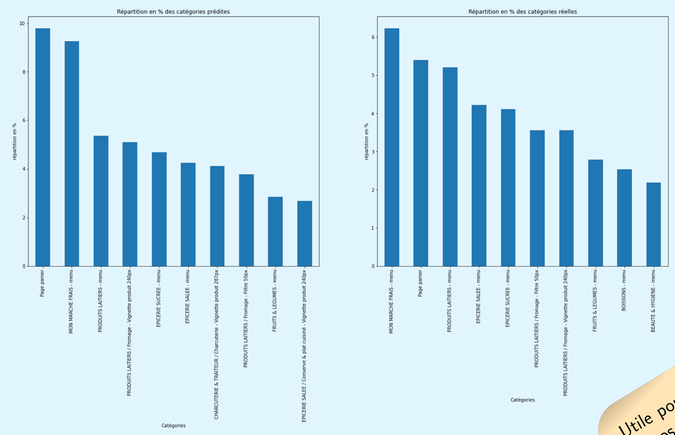
Prétraitement des données



Lucky Cart dispose d'une grande base de données sur les clients d'Intermaché, elle contient plus de 130 Giga octets de données. Ainsi, lorsque nous récupérons ces données, il est nécessaire de les nettoyer, car les données ne sont pas toujours dans le format voulu.

Le prétraitement des données est une étape essentielle lors du commencement d'un projet d'apprentissage machine, sans cela, nos données sont inutilisables et non interprétables par un ordinateur. Ainsi, ce processus a été l'étape la plus importante du projet, car construire des modèles sur des données mauvaises, implique que le modèle sera mauvais.

Visualisation des prédictions



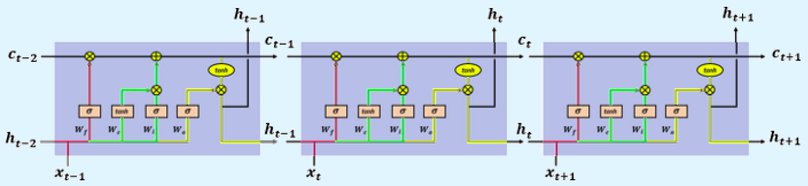
Utile pour avoir une idée de nos prédiction

Évaluation des modèles

Notre meilleur modèle : Le LSTM sans plongement des catégories.

MODÈLES			Top K précision catégorique K=1	Top K précision catégorique K=3
Feed Forwards	Hyperparamètres non-optimisés	Ø	35.84%	54.27%
LSTM		Ø	36.81%	55.92%
LSTM+Plongement		Ø	36.38%	55.21%
LSTM	Hyperparamètres optimisés	Hyperband	Coming soon	
		Bayésien		
LSTM+Plongement		Hyperband	36.78%	55.88%
		Bayésien	36.58%	55.49%

Notre modèle : le réseau LSTM



Le LSTM est composé de 3 "portes" qui mettent à jour et contrôle l'état du neurone:
La **porte d'oublie** contrôle les informations que le neurone doit "oublier" en fonction des nouvelles informations reçues. La **porte d'entrée** contrôle quelles nouvelles informations seront encodées dans le neurone en fonction des nouvelles informations reçues. La **porte de sortie** contrôle quelles informations du neurone seront envoyées comme entrée dans le prochain pas de temps. Cette configuration permet au LSTM de pallier l'effet de "Disparition du gradient" que l'on retrouve dans les réseaux de neurones récurrents classiques.

Problèmes majeurs

- Le manque de ressources de traitements pour l'optimisation des hyperparamètres, notamment le manque de RAM nous a empêché d'optimiser les hyperparamètres de manière efficace. De plus, le problème lié au CPU de nos machines pas assez performantes pour du calcul scientifique ne nous a pas permis d'améliorer autant qu'on l'espérait notre réseau LSTM avec plongement des catégories. Si des machines destinées aux calculs scientifiques nous avaient été mises à disposition, peut-être aurions-nous dépassé les 40% de précision catégorique ?