MACHINE LEARNING

Master 2 - Data & IA - FGES

COMPETENCES EVALUEES

- Analyser le corpus de données disponibles afin de choisir les données appropriées, réduire l'espace de l'IA.

- Structurer les données dans une base de données, dans le respect de la réglementation en termes de protection des données individuelles (RGPD) et en collaboration avec le DPO en vue de produire un prototype afin de les exploiter dans la solution I.A.

- Évaluer l'adéquation des modèles d'apprentissages disponibles avec le projet IA et la solution proposée (réseau neuronaux, arbre de décision, forêt aléatoire, boosting, clustering, etc.).

- Concevoir un modèle (non existant) ou adapter un modèle (existant) à partir des spécificités des données de l'entreprise en utilisant les analyses statistiques et mathématiques afin de répondre à la solution I.A proposée.

- Proposer des modèles statistiques et de data science (machine learning) à mettre en pratique aux directions métiers afin de détecter des nouveaux services, anticiper des besoins et résoudre des problématiques métiers de l'entreprise.

- Analyser les performances et la capacité prédictive d'un modèle proposé dans la solution I.A.

- Définir une procédure d'entraînement adéquate d'un modèle en sélectionnant des données d'apprentissage les plus adéquates au besoin d'analyse du projet IA afin de la mettre en place.

- Définir une phase de test et de validation du modèle d'apprentissage choisi afin de le mettre en place.

- Maquetter l'infrastructure nécessaire à la mise en place de la solution IA afin de permettre la réalisation de son déploiement et de son fonctionnement par les équipes projet.

CONTEXTE DU PROJET

Amazing est une marketplace en ligne qui propose une grande variété de produits. C'est un leader sur le marché mondial. Une partie de son chiffre d'affaires est engendrée par sa marque propriétaire : Amazing Basics. Elle propose un grand choix de produits dans des catégories très variées (technologies, prêt-à-porter, accessoires de maison, etc.).

Depuis la dernière inflation, Amazing fait face à une baisse du chiffre d'affaires sur ses produits Amazing Basics, notamment sur les biens de divertissement. Lors de différentes conférences dans le domaine de la technologie, Amazing a beaucoup entendu parler de la pertinence d'utiliser des modèles d'IA ou de machine learning pour booster les ventes. Cela prenait différentes formes : recommandation de produits, prédiction des ventes, modélisation de clients-types...

Afin de rester au niveau de ses concurrents, Amazing fait appel à une équipe de data scientists pour mener un projet leur permettant de booster leurs ventes à moindre coût et en surfant sur la vague de l'IA. Amazing aimerait notamment mieux connaître ses clients, pour adapter son offre et ses prix ou personnaliser l'expérience d'achat. Votre expertise en analyse de données et en machine learning est cruciale pour aider Amazing rester à la pointe de l'innovation et booster ses ventes en ligne.

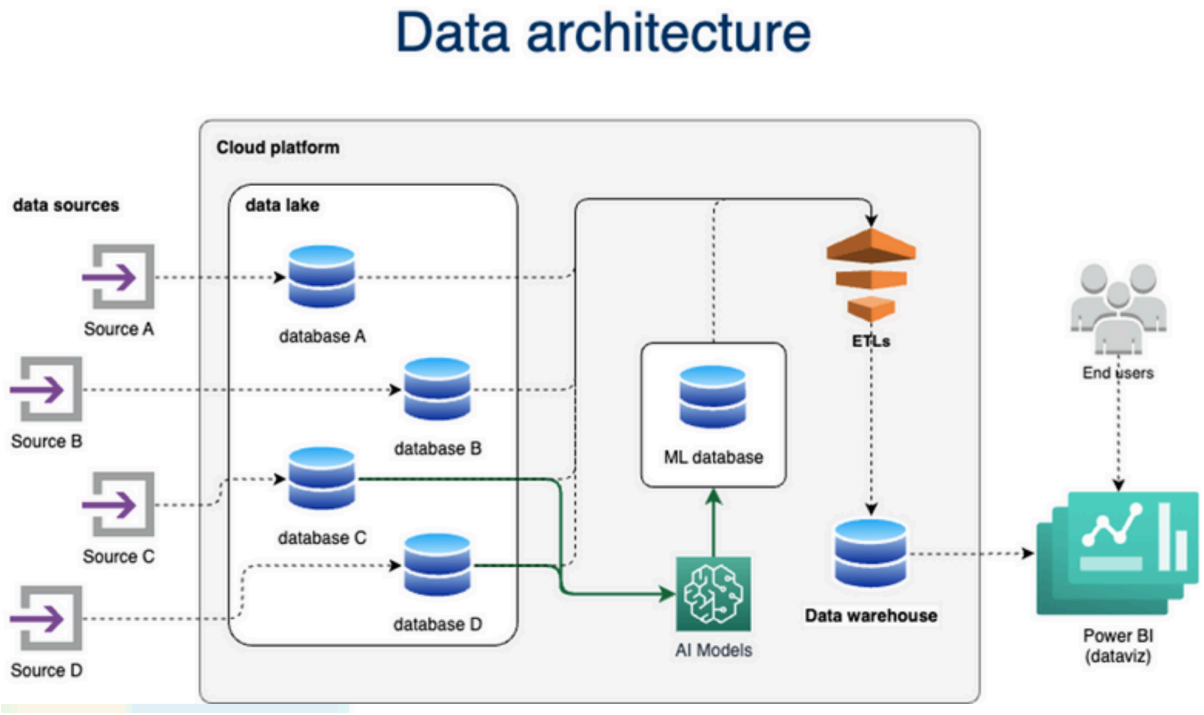
Pour ce projet, vous travaillerez en collaboration avec l'équipe Marketing et Business Intelligence de la société. Ce sont eux qui sont à l'origine du projet et qui utiliseront ses résultats. Vous pourrez donc vous appuyer sur les ressources internes d'Amazing pour proposer des actions à mener grâce aux résultats de votre projet (envoi de newsletters, la mise en place de promotions spécifiques, etc.).

Parmi ses équipes, Amazing dispose de deux Data Engineers qui vous fournissent les données nécessaires et peuvent mettre en place une architecture cloud ou on-premise pour industrialiser le projet. L'architecture data d'Amazing est hébergée sur le cloud en un data lake.

Les data engineers de Amazing peuvent vous fournir les données suivantes :

- Catalogue des produits (Nom, prix, catégorie, ..)
- Commandes effectuées (avec les informations du client, des produits, de la commande)
- Comportement des utilisateurs sur le site (visite de page, clic, durée de session)
- Base de données des clients anonymisée
- Base de données de gestion des stocks
- Historique des actions marketing : newsletters, promotions, calendrier des fêtes (black friday, cyber monday, etc.)

Le schéma ci-dessous représente l'architecture data présente chez Amazing. Votre projet s'intégrera dans les éléments en vert (AI Models).



SPÉCIFICATION DU BESOIN

Le modèle réalisé devra permettre de catégoriser la base clients d'Amazing. Il doit être applicable sur n'importe quel client (actuel ou futur) dès lors qu'il a réalisé un certain nombre d'évènements sur le site. (Un évènement peut être : la visite d'une page produit, l'ajout au panier, le retrait d'un article du panier, ou un achat.) Pour rappel, Amazing ne cherche pas à catégoriser ses clients par caractéristiques démographiques ou sociales **mais bien par leurs habitudes d'achat et de visites sur le site**.

Chaque catégorie devra faire l'objet d'une analyse pour en extraire ses caractéristiques principales. On cherche à en comprendre ce qui fait sa singularité pour comprendre les clients qui la composent. Généralement, un nom est attribué à chaque groupe pour comprendre en un coup d'œil le type de client dont il s'agit.

Modèle

Le modèle pourra prendre en compte diverses dimensions, telles que le type de produit, la fréquence d'achat, le montant dépensé, les préférences saisonnières, etc. Différents algorithmes pourront être envisagés (K-NN, Decision tree, SVM).

La réalisation du modèle pourra faire l'objet d'une analyse des importances des features et leur sélection.

Une analyse des composantes principales peut s'avérer pertinente pour prendre en compte de façon plus efficace un grand nombre de features.

Données

Pour répondre à cette problématique, Amazing met à disposition une extraction de données d'événements réalisés sur son site entre octobre 2019 et avril 2020.
L'équipe de Data Scientist devra transformer le jeu de données en jeu de caractéristiques pour chaque utilisateur (user_id) comportant un ensemble de métriques pertinentes au modèle. Il est laissé libre à l'équipe de Data Scientists de définir, en justifiant, le seuil optimal (en nombre d'évènements) à partir duquel il est fiable de catégoriser un client.

LIVRABLES ATTENDUS

- Préparer un dossier présentant votre démarche, les résultats obtenus. Ce dossier sera à remettre en PDF et fera l'objet d'une présentation orale par groupe incluant une démonstration technique.
- Définir les métriques pertinentes à calculer pour qualifier un utilisateur (variables explicatives).
- Intégration des données dans une base de données (relationnelle ou non relationnelle)
- Réaliser une analyse descriptive sur les données à disposition. Préparer un nettoyage si nécessaire
- Mettre en place le traitement des données (nettoyage, calcul des variables explicatives).
- Le modèle doit être capable de traiter de nouveaux fichiers d'événements au cours du temps (au même format que ceux fournis).
- Concevoir un ou plusieurs modèles répondant à la problématique. Le/les optimiser en analysant les performances et résultats.
- Réaliser une exploration des catégories finales afin d'établir un compte rendu de ce qui caractérise chacune d'elles.

LIEN VERS LES DONNEES

Le jeu de données est disponible aux liens suivants :

2019-Oct.csv.gz (1.62Gb) [https://data.rees46.com/datasets/marketplace/2019-Oct.csv.gz]
2019-Nov.csv.gz (2.69Gb) [https://data.rees46.com/datasets/marketplace/2019-Nov.csv.gz]
2019-Dec.csv.gz (2.74Gb) [https://data.rees46.com/datasets/marketplace/2019-Dec.csv.gz]
2020-Jan.csv.gz (2.23Gb) [https://data.rees46.com/datasets/marketplace/2020-Jan.csv.gz]
2020-Feb.csv.gz (2.19Gb) [https://data.rees46.com/datasets/marketplace/2020-Feb.csv.gz]
2020-Mar.csv.gz (2.25Gb) [https://data.rees46.com/datasets/marketplace/2020-Mar.csv.gz]
2020-Apr.csv.gz (2.73Gb) [https://data.rees46.com/datasets/marketplace/2020-Apr.csv.gz]

Feature	Description
event_time	Time when event happened at (in UTC).
event_type	Type of event
product_id	ID of a product
category_id	Product's category ID
category_code	Product's category taxonomy (code name) if is was possible to make it. Usually present for meaningful categories and skipped for different kinds of accessories.
brand	Downcased string of brand name. Can be missed
price	Float price of the product.
user_id	Permanent user ID.
user_session	Temporary user's session ID. Same for each user's session. Is changed every time user come back to online store from a long pause

PLANNING

Début: 01 octobre 2024
Fin: 09 décembre 2024

Durée totale: 26 heures en tutorat avec votre enseignant par groupe.

Modalités d'évaluation:

- Livrables
- Présentation des différents groupes (30 minutes) + questions / réponses:
    - 16 décembre 2024 (3 groupes)
    - 07 janvier 2024 (3 groupes)