Mise en situation de projet réel : Modèle IA

Formation : 2ème année Mastère 2 Data Analyst, Mastère 2 Business intelligence et Data science, Mastère 2 Prévention et analyse des risques

Année : 2024-2025

Titre RNCP : 36129 – Chef de Projet Intelligence Artificielle

Bloc 3 : Développer une solution d'intelligence artificielle (Machine et Deep learning)

Type d'épreuve : Dossier écrit

Date de remise : 11/04/2025 à 12h00 dernier délai

CONCEVOIR UNE SOLUTION IA

Cahier des Charges de la MSPR « Conception d'un modèle IA »

COMPÉTENCES ÉVALUÉES :

C15 - Concevoir un modèle IA en élaborant le Design de l'architecture informatique de la solution IA à développer via une « Application Programming Interfaces » (API), en définissant des objectifs de performance visés, en sélectionnant un ou plusieurs algorithmes adaptés au projet d'apprentissage automatisé envisagé, supervisé ou non supervisé (supervised / unsupervised learning), afin d'exploiter les résultats du prototypage.

C16 - Optimiser le modèle IA en interprétant les premiers résultats obtenus en contrôlant la qualité des modèles prédictifs – Time-series Predictions / Predictive Analytics – à l'aide de scénarios de test préétablis – tests théoriques ou cas d'usage réels, en analysant la fiabilité de l'algorithme par rapport au niveau de performance ou de précision attendu, afin d'améliorer l'algorithme à partir des évaluations réalisées.

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.

Résultats attendus :

- Une solution d'apprentissage profond est proposée, opérationnelle et en adéquation avec la problématique métier

- Des indicateurs permettant de mesurer les performances des modèles d'apprentissage sont proposés

- L'algorithme d'intelligence artificielle sélectionné est adapté à la problématique métier- L'implémentation de la solution dans l'environnement informatique est réussie

- Le choix des méthodes d'échantillonnage et des métriques d'évaluation des algorithmes d'intelligence artificielle sont expliqués

- La proposition d'optimisation des algorithmes permet d'améliorer l'efficacité des modèles

d'apprentissage automatisé

- Une stratégie de validation croisée des données est proposée et permet de minimiser l'influence des valeurs extrêmes.

Le dossier devra contenir l'ensemble des éléments demandés, en particulier des plans d'actions présentant à une Direction générale les principes adoptés et les principaux parcours usagers.

Il est attendu un dossier écrit de 10 pages hors annexe à remettre le 11/04/2025 à 12h00 dernier délai.

1 - CONTEXTE

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 un(e) data scientist 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.

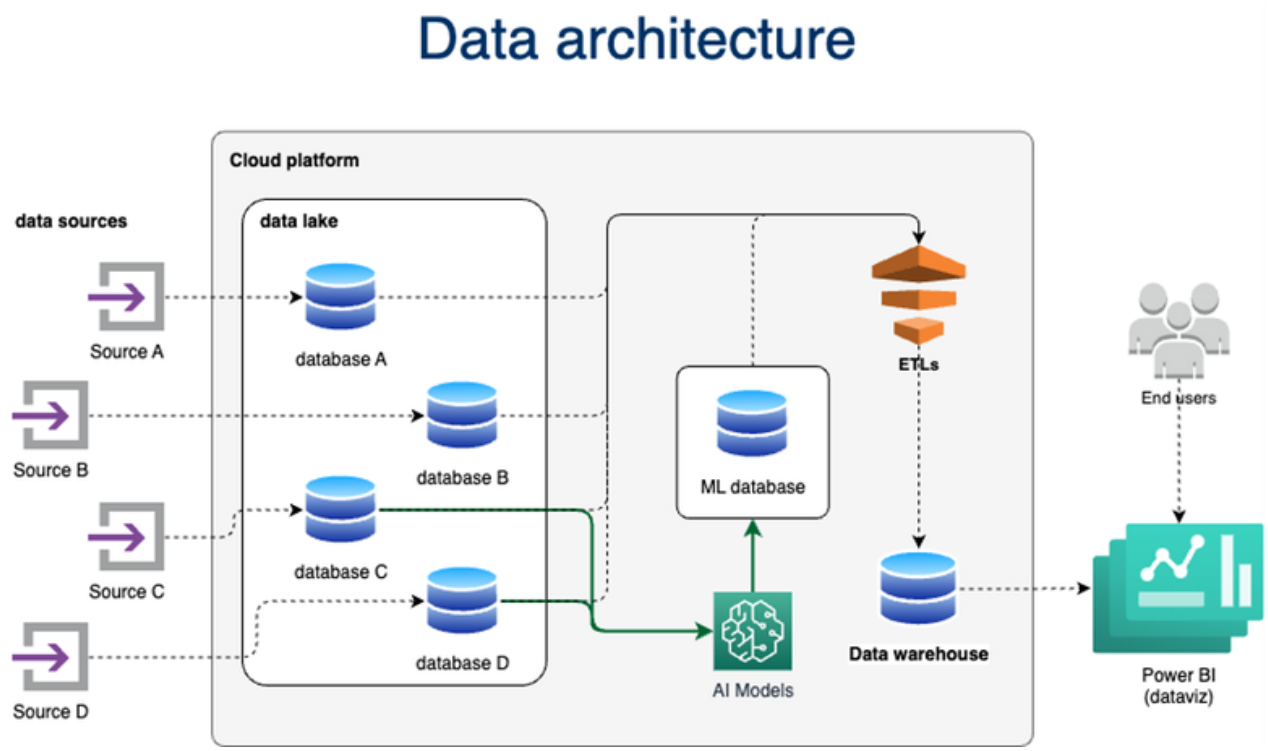
Pour ce projet, vous **travaillerez seul** mais en collaboration avec l'équipe Marketing et Business Intelligence de la société. Ce sont eux qui sont à l'origine du projet et qui utiliseront ses résultats. Vous pourrez donc vous appuyer sur les ressources internes d'Amazing pour proposer des actions 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'intègrera dans les éléments en vert (AI Models).



2 – SPÉCIFICATIONS 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.

2.1 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. Il est important de noter que ce modèle et son ETL doivent être capable de fonctionner sur de gros volumes de données au moment de son industrialisation.

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

2.2 Données

Pour répondre à cette problématique, Amazing met à disposition une base de données d'évènements réalisés sur son site entre octobre 2019 et avril 2020.

En tant que Data Scientist vous devrez transformer le jeu de données en jeu de caractéristiques pour chaque utilisateur (user_id) comportant un ensemble de métriques pertinentes au modèle.

Vous avez la possibilité de définir, en justifiant, le seuil optimal (en nombre d'évènements) à partir duquel il est fiable de catégoriser un client.

2.3 Industrialisation

L'industrialisation du modèle sera prise en charge par l'équipe data de Amazing. Cependant, pour s'intégrer à l'architecture technique déjà existante, l'algorithme de classification doit être contenu dans un container type Docker ou Kubernetes et doit être documenté. Les résultats de la classification seront stockés dans une base de données Redshift, accessible au Data Warehouse.

3 – LIVRABLES ATTENDUS

- Définir les métriques pertinentes à calculer pour qualifier un utilisateur (variables explicatives).
- 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) à l'aide d'un ETL. Le modèle doit être capable de traiter de nouveaux fichiers d'évènements au cours du temps (au même format que ceux fournis).
- Mettre en place une méthodologie en conformité avec le RGPD.
- 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.
- Concevoir l'architecture nécessaire pour l'industrialisation du modèle.

Le dossier devra contenir l'ensemble des éléments demandés, en particulier des plans d'actions présentant à une Direction générale les principes adoptés et les principaux parcours usagers.

LES COMPETENCES EVALUEES DURANT CETTE MSPR :

CRITERES	PONDERATION
-Solution d'apprentissage en adéquation avec la problématique	3
- Des indicateurs de performances des modèles d'apprentissage	3
- L'algorithme d'intelligence artificielle sélectionné adapté à la problématique métier	2
- L'implémentation de la solution dans l'environnement informatique	3
- Les choix des méthodes d'échantillonnage et des métriques d'évaluation des algorithmes d'intelligence artificielle sont expliqués	3
- La proposition d'optimisation pour améliorer l'efficacité des modèles	2
- Une stratégie de validation croisée des données qui permet de minimiser l'influence des valeurs extrêmes.	3

ANNEXE – DONNÉES

Le jeu de données est disponible aux liens suivants :

https://drive.google.com/drive/folders/10toga2qvo7ISxuaCQGSXy2aZVauzrbWe?usp=drive_link

Amazing y a entreposé les éléments suivants :

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.