DATASET 1

Use Case 1

Problème métier

- Il faut pouvoir estimer les ventes journalières à l'avance afin d'adapter les stocks en conséquence et éviter les ruptures ou le surplus de stock.
- Si chaque magasin estime ses ventes dans son coin, on pourrait avoir des méthodes d'estimation différentes et donc une difficulté pour comparer et avoir une vision globale.

Solution technique

- Estimer les ventes journalières à l'avance en utilisant les données de ventes de tous les magasins et les informations telles que le jour de la semaine, le fait qu'il y ait des promotions ou des vacances à partir d'un modèle de Machine Learning.
- L'évaluation du modèle se fait à partir de la différence entre les ventes prédites et les ventes réelles.

RH + temps + coût financier

- 1 data analyst et 2 data scientists à temps plein pour pour traiter les données et développer le modèle pendant 1 mois (1m).
- 2 développeurs pendant un mois (1m) pour le déployer.
- 1 manger pour suivre le projet
- Total : 35 000€ pour 2 mois

INSIGHTS

IA

- Features X: DayOfWeek, Customers, Promo, StateHoliday, SchoolHoliday
- Target variable y: Sales
- Model : Regression model (xgbRegressor)
- Score $= |y_{true} - y_{pred}|$

$$= \sum_{i=1}^{n} |y_{true}^{i} - y_{pred}^{i}|$$

Avantages financiers

- Suppression du coût d'estimation du CA pour chaque magasin (RH + temps).
- Réduction des coûts de stockage via l'optimisation des stocks.
- Augmentation des bénéfices via l'évitement de la rupture de stock.

DATASET 1

Use Case 2

Problème métier

- Pour mettre en place une politique de vente, il faut pouvoir prendre en compte les spécificités de chaque magasin.
- En effet, les jours d'ouvertures, période de vacances, de promotion et le nombre de clients pouvant être différents, il faut adapter la stratégie en fonction.
- Certains magasins peuvent tout de même partager des similarités.

Solution technique

- Regrouper les magasins ayant un profil similaire en appliquant un algorithme de Machine Learning non supervisé.
- Ainsi, les magasins sont regroupés dans des classes. Les magasins qui sont dans la même classe sont le plus semblables possibles (variance intra-classe minimum) tandis que les classes sont le plus dissemblables possibles entre elles (variance inter-classe maximale) .

RH + temps + coût financier

- 1 data engineer pour traiter les données pendant 2 semaines (2s) et 1 data scientist à temps plein (1m) pour concevoir le modèle.
- 1 développeur pour le déployer (2s)
- 1 manager pour suivre le projet
- Total : 20 000 € pour 2 mois

INSIGHTS

IA

- Features X: DayOfWeek, Customers, Promo, StateHoliday, SchoolHoliday
- Réduction de dimension via ACP
- Algorithme : CAH (Classification Ascendante Hiérarchisée)
- Métrique de distance : distance de Gower

Avantages financiers

- Coût de marketing réduit puisque restreint à quelques profils pour établir la stratégie
- Outil d'aide à l'établissement de la stratégie marketing et amélioration de son efficacité.
- Gain apportée par la détection potentielle d'anomalies sur certains magasins et visualisation des facteurs qui permettent à d'autres magasins d'avoir plus de succès

DATASET 2

Use Case 1

Problème métier

- Le dataset contient beaucoup de valeurs manquantes, notamment sur certaines stations et lignes de production.
- Cela nuit à la continuité de l'information et à la compréhension du fonctionnement de chaque ligne de production
- On a un problème de type Missing Not At Random (MNAR) où certaines stations ont davantage de valeurs manquantes.

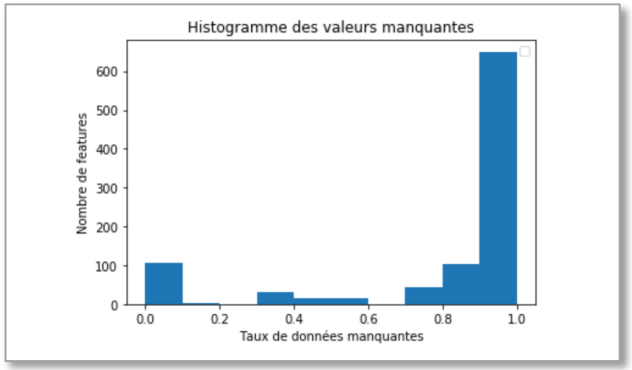
Solution technique

- Remplacer les valeurs manquantes par des valeurs plausibles au vu des données récoltées sur la ligne de production mais également sur les autres stations et lignes de production.
- Le remplacement des valeurs ne se fera pas avec une simple imputation par la moyenne mais avec un imputation multiple à partir de simulations prenant en compte la variabilité des données.

RH + temps + coût financier

- 1 data analyst pour regrouper les données (timestamp+mesures) (2s) et un data scientist pour concevoir le modèle d'imputation (1m et demi).
- 1 expert du domaine de production pour expliquer les données et superviser la cohérence des résultats.
- Coût total : 15 000 € pour 2 mois.

INSIGHTS



IA

- Données : timestamps et mesures
- Générer M datasets imputés
- Algorithme : Bootstrap EM

❶ Bootstrap rows: $X^1, \ldots, X^M$
EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1), \ldots, (\hat{\mu}^M, \hat{\Sigma}^M)$

❷ Imputation: x_{ij}^m drawn from $\mathcal{N}\left(\hat{\mu}^m, \hat{\Sigma}^m\right)$

Avantages financiers

- Les nouvelles données pourront être utilisées pour les autres cas d'usage et permettront d'améliorer la performance des modèles.
- L'homogénéisation entre les différentes lignes de production apportera davantage d'informations pour le suivi de la production, ce qui éclairera le choix réalisés (ouverture de nouvelles lignes par exemple).

DATASET 2

Use Case 2

Problème métier

- Un des gros challenges auquel doivent faire face les industriels est le respect des normes de production, qui sont de plus en plus strictes.
- Un défaut sur un produit peut causer le rappel de millier d'exemplaires et entrainer un coût très important à l'arrivée.

Solution technique

- Un modèle permettra d'indiquer en fonction des différentes mesures réalisées sur une ligne de production s'il y a un risque d'échec des tests de conformité.
- Il s'agira donc d'un modèle de classification en ligne qui traitera un flux de données (entrainé sur une fenêtre de temps) en continu et détectera les risques de défaut qu'il pourra prévenir via un signalement.

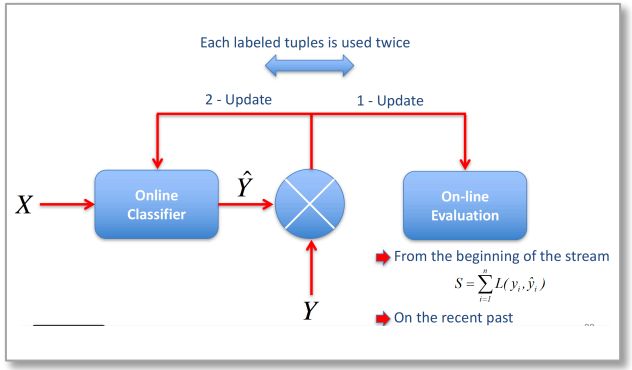
RH + temps + coût financier

- 1 data engineer pour concevoir l'infrastructure de donnée gérant les data stream (2m).
- 1 Business analyst pour récupérer les données des test de conformité (1m).
- 1 data analyst pour assembler les différentes données (timestamps+mesures+tests) (1m)
- 1 data scientist pour concevoir le modèle d'apprentissage en ligne (2s).
- 2 développeurs pour intégrer un signalement dans les plateformes internes (1m ½).
- Coût total : 50 000 € pour 6 mois

Conditions

- Pour pouvoir mettre en place un modèle prédictif de conformité, il faut pouvoir récupérer l'historique des tests de conformité en plus des mesures réalisées afin de modéliser le lien entre les deux.
- Les usines tournant à plein temps, il faudrait aussi pouvoir traiter un flux important de données en continu et non pas se baser sur une base de données statiques.

IA



Avantages financiers

- Prévenir le risque de défaut / non-respect des normes d'un produit permettra d'éviter les rappels de produit dont le coût peut être très élevé (sans compter le risque judiciaire).
- Cela permettra également d'éviter les arrêts de production brutaux et de réduire les coût de maintenance à fortiori ainsi que la satisfaction des client fournis.

DATASET 3

Use Case 1

Problème métier

- La base contient des informations sur les clients d'une banque, notamment sur le type de produit qu'ils détiennent (prêts, dépôts, titres, etc.).
- Or, la plupart des clients ne sont pas intéressés par certains produits comme les titres, il serait donc préférable de cibler les campagnes marketing sur ceux qui seraient susceptibles d'être intéressés.

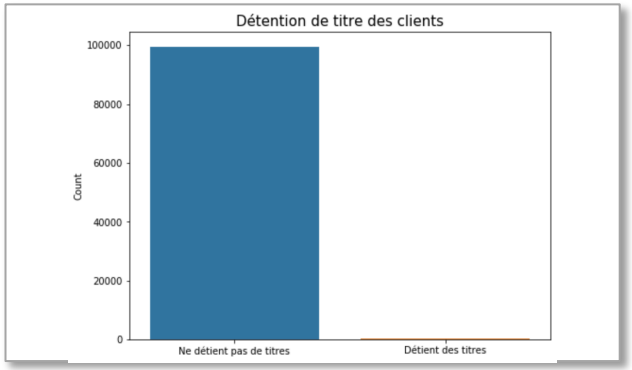
Solution technique

- Un modèle indiquera pour chaque produit si le client est susceptible de s'y intéresser. Le modèle sera basé sur les choix des clients par le passé pour modéliser leur choix futur (un client qui détenait tel produit a choisi tel autre).
- On aura donc plusieurs modèles de classification binaire (un par produit) qui renverra 0 si le client n'est probablement pas intéressé ou 1 s'il est fortement susceptible de l'être.

RH + temps + coût financier

- 2 data scientist pour travailler sur les modèles de classification (1m)
- 3 développeurs pour intégrer chaque modèle dans une plateforme unique (2m)
- 1 expert marketing pour suivre le projet
- Total : 30 000 € pour 3 mois

Insights



IA

- Données : données sur les clients et leurs produits.
- Target : acquisition de chaque produit (0 ou 1).
- Modèle : **Forêt aléatoire** afin d'obtenir des règles interprétable et endiguer le problème de déséquilibre des données (entre les 2 classes) grâce à l'échantillonnage (Bootstrap).

Avantages financiers

- Cibler les clients susceptibles d'être intéressés permettra de réduire les coûts de marketing.
- Un marketing personnalisé pour ces clients augmentera les chances de les séduire.
- Ne pas afficher de la même publicité en masse pour tous les clients permettra d'améliorer l'image de la banque, plus proche de ses clients.

DATASET 3

Use Case 2

Problème métier

- Les produits financiers n'étant pas toujours compris par tout le monde, les clients peuvent être réticents à élargir leur portefeuille de produits.
- Certains clients pourraient ne pas avoir l'intention d'acheter un certain produit (ce que pourrait nous pointer le modèle précédent) alors que celui-ci est recommandable.

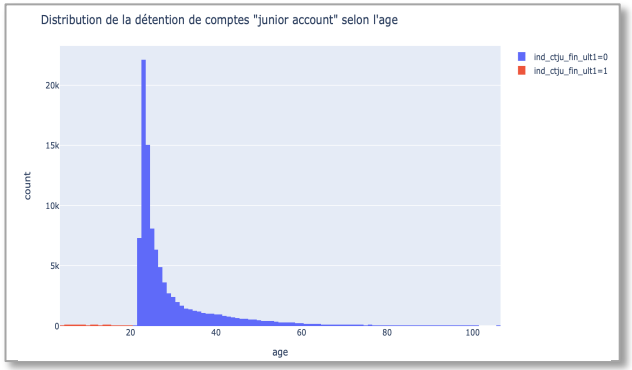
Solution technique

- Le modèle recommandera 3 produits (top-3) selon leur un score qui leur est attribué.
- Il sera basé sur le collaborative filtering : les personnes similaires sont susceptibles d'avoir les mêmes préférences.
- Les recommandations sont donc basées sur les choix de produits des individus (l'avantage étant que les informations sur les produits ne sont pas nécessaires).

RH + temps + coût financier

- 1 data scientist pour travailler sur le système de recommendation (2 semaines)
- 2 développeurs pour concevoir l'application et l'intégrer sur la plateforme de la banque (1m et demi)
- 1 manager pour suivre le projet
- Coût total : 15 000 € pour 2 mois

Insights



IA

- Données : données sur les clients et leurs produits
- Algorithme : k-Nearest Neighbor (k-NN)
  On détermine les « voisins » du client à partir d'une mesure de similarité et on détermine le score pour chaque produit en fonction des choix du voisin.

$$\hat{r}_{ui} = \frac{\sum_{v \in \mathcal{B}(u)} sim(u,v) \cdot r_{vi}}{\sum_{v \in \mathcal{B}(u)} sim(u,v)}$$

Avantages financiers

- Une bonne recommandation des produits permet de diversifier le portefeuille des clients.
- Un système de recommandation est un bon moyen de fidéliser les clients
- Un système de recommandation peut permettre d'attirer de nouveaux clients en leur proposant les offres adaptées à leurs besoins.

DATASET 4

Use Case 1

Problème métier

- Lorsqu'un client a a réalisé une commande, on ne sait pas quand est-ce qu'il va revenir.
  En effet, les fréquences d'achat varie fortement d'un client à un autre selon qu'il soit membre d'une famille nombreuse ou non (beaucoup de client dans les 2 extrêmes).
- Or, avoir une idée de la date de la prochaine commande de chaque client permettrait d'ajuster l'offre et la stratégie marketing pour les proposer au bon moment.

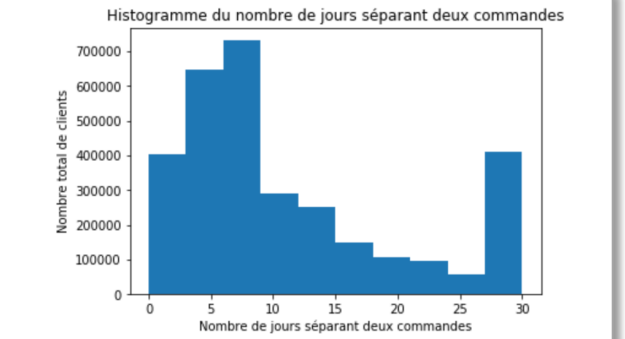
Solution technique

- Le modèle permettra de donner pour chaque client la durée qui va séparer sa dernière commande de sa prochaine commande.
- Il s'agira d'un modèle de régression qui modélisera cette durée en fonction des produits que le client achète et ses heures d'achat, ainsi que les durées qui ont séparées deux commandes du client par le passé.

RH + temps + coût financier

- 1 Business analyst pour récupérer les dates de chaque commande + 1 data analyst pour assembler les différentes tables (1m)
- 1 data scientist pour concevoir le modèle (2s) et 1 développeur pour le déployer (2s)
- 1 manager pour suivre le projet
- Total : 20 000 € pour 2 mois

Insights



IA

- Features : données sur les commandes des clients (heure, durée entre 2 commandes, produits achetés, …).
- Target : Durée (en jours) séparant la dernière commande de la prochaine commande.
- Algorithme : XGBoost Regressor
- Score = $|y_{true} - y_{pred}|$
- Donnée à ajouter : Date de chaque commande (pour pouvoir utiliser le modèle en pratique).

Avantages financiers

- Connaître les clients qui ont une faible fréquence d'achat permettrait de les cibler en termes de marketing pour les attirer avec des promotions.
- Savoir quand est-ce que les clients vont revenir permettra d'assurer qu'ils retrouvent leurs produits préférés (éviter les ruptures de stock) et ainsi augmenter la satisfaction client.

DATASET 4

Use Case 2

Problème métier

- Les habitudes d'achat des clients sont dynamiques. En effet, un client peut adapter ses comportements d'achat selon les tendances du moment (acheter bio par exemple).
- Ainsi, une fois que l'on sait (à peu près) quand est-ce qu'un client va revenir, il serait bon de savoir quels produits il est susceptible d'acheter et supposer qu'il va acheter les mêmes produits que la dernière fois serait tout du moins présomptueux.

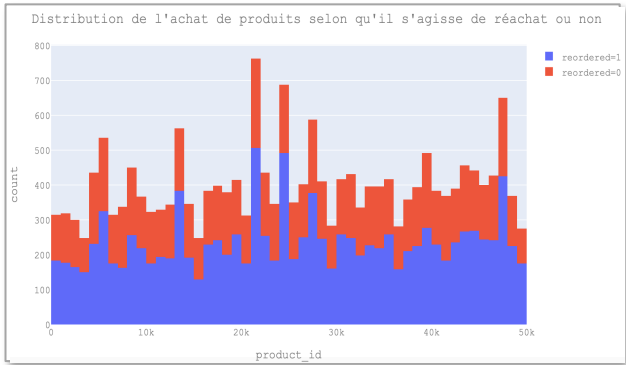
Solution technique

- Le modèle permettra de savoir pour chaque produit qui a été acheté par le client durant sa dernière commande, s'il sera acheté à nouveau lors de sa prochaine commande.
- On aura plusieurs modèles de classification binaire (un par produit acheté) qui renverra 0 si le client ne va probablement pas le racheter ou 1 s'il y a de fortes chances qu'il rachète le produit.

RH + temps + coût financier

- 1 data analyst pour assembler les différentes tables (1m)
- 2 data scientists pour concevoir le modèle (1m)
- 1 développeur pour le déployer (1m)
- 1 manager pour suivre le projet
- Total : 30 000 € pour 3 mois

Insights



IA

- Données : données sur les commandes des clients et les produits achetés
- Target : réachat de chaque produit (0 ou 1)
- Modèle : XGBoost Classifieur (modèle le plus performant pour ce type de problème).
- Score : F1 score

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Avantages financiers

- Connaître en avance les produits qui ont de fortes chance d'être achetés par le client lors de sa prochaine visite permettra de prévenir les ruptures de stock et assurer qu'il trouvera bien ce qu'il cherche en rayon.
- Si le client trouve à chaque ses produits, il sera plus satisfait et cela permettra de le fidéliser.