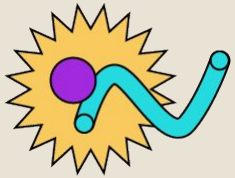
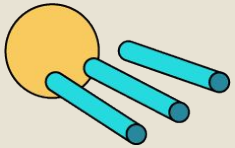
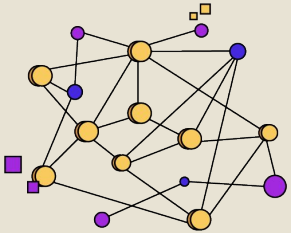


Test Technique de Data Science

DataScientest

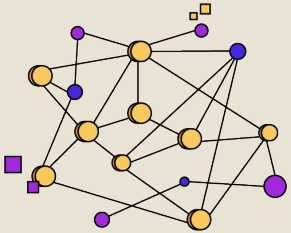




Déroulé d'un test technique DS

Déroulé d'un test technique DS

1. Programmation sur Python
   - code et algorithmes sur Python
   - focalisation sur les librairies Python de Data Science
2. Use Case de Machine Learning
   - compréhension métier
   - méthodologie et détails des Étapes de résolution
   - vulgarisation des résultats d'un point de vue business
3. Questions techniques générales



Partie 1 : Programmation sur Python

Programmation sur Python

Code et algorithmes sur Python :

1. Ecrivez une fonction qui prend en entrée deux chaînes de caractères et qui renvoie si elles sont des anagrammes ou non.
2. Ecrivez une fonction qui renvoie la factorielle d'un nombre.

Pour rappels, la factorielle d'un nombre est le produit des nombres entiers inférieurs ou égaux à ce nombre.

3. Ecrivez un code permettant de renvoyer la pyramide de nombres suivante :

```
1

2 3

4 5 6

7 8 9 10
```

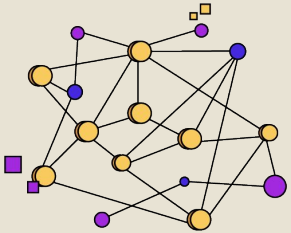
Programmation sur Python

Python en Data Science

1. Importez les librairies Python les plus utilisées pour un projet de Data Science.
2. Chargez le fichier `titanic.csv` sous forme de dataframe Pandas.
3. Extrayez les informations des passagères de plus de 30 ans.
4. Classez les passagers en fonction de leur classe et de leur âge.
5. Affichez la proportion de survivants par classe.
6. La passagère de 48 ans en classe 2 qui a survécu n'a pas de valeur renseignée pour sa cabine. Remplacez la valeur manquante par la modalité "C85".
7. La variable "Fare" renseigne sur le prix du ticket. Affichez les recettes totales du Titanic par classe.
8. Affichez les distributions des variables "Sex" et "Age" sous forme de plots.

Programmation sur Python

- savoir définir une fonction Python (et notamment les fonctions usuelles)
- connaître les méthodes existantes pour les listes et arrays
- maîtriser l'utilisation des boucles (notamment les boucles imbriquées)
- connaître les principales méthodes Pandas pour les dataframes
- penser à l'utilisation de pd.value_counts(), pd.sort_values(), pd.groupby(), pd.agg(), pd.apply()



Partie 2 : Use Case de Machine Learning

Use Case de Machine Learning

Selon le Harvard Business Review, un indicateur important pour les entreprises dont les bénéfices sont basés sur des paiements récurrents est le “rate of churn”. Le “rate of churn” est le taux de désabonnement des clients, qui se tournent alors vers une entreprise concurrente.
Dans votre cas, vous travaillez pour une banque qui souhaite connaître ses clients susceptibles de partir, afin d’améliorer leur expérience et limiter les départs. L’objectif final est donc de diminuer le “rate of churn”.
En tant que Data Scientist, de quelles données auriez-vous besoin et quel serait votre processus de résolution pour aider la banque en question ?

Use Case de Machine Learning

1. Se focaliser sur le contexte métier pour déterminer les données utiles et ne pas hésiter à poser des questions supplémentaires

- problématique : quelle stratégie adopter pour diminuer le nombre de clients qui changent de banque ?
- problématique data : comment prédire les clients qui vont quitter la banque et quels facteurs impactent cette prédiction ?

⇨ churn prediction avec du Machine Learning

- données nécessaires : données clients avec leurs caractéristiques individuelles et leurs caractéristiques bancaires

⇨ possible de récupérer les données via la banque (potentiellement anonymisées)

Use Case de Machine Learning

2. Présenter son code en expliquant le raisonnement suivi et les Étapes clés de résolution

Etape 1 : Exploration de Données

CustomerId : identifiant unique du client
Surname : nom du client
CreditScore : note attribuée au client en fonction de ses emprunts bancaires (un CreditScore élevé est associé à une note élevée)
Geography : localisation du client
Gender : genre du client
Age : âge du client
Tenure : nombre d'années depuis lesquelles le client appartient à cette banque
Balance : solde du client sur son compte bancaire
NumOfProducts : nombre de produits auxquels le client a souscrits chez cette banque
HasCrCard : variable binaire indiquant si le client possède une carte de crédit chez cette banque
IsActiveMember : variable binaire indiquant si le client est un client actif de cette banque
EstimatedSalary : salaire estimé du client
Exited : variable binaire indiquant si le client a quitté la banque (0 : il n'a pas quitté, 1 : il a quitté)

Use Case de Machine Learning

Etape 2 : Analyse de Données avec la DataVizualization

- Distribution de la variable cible (univarié)
- Analyse descriptive des variables (univarié)
- Analyse multivariée : impact des variables explicatives sur la variable cible (bivarié) et corrélations des variables explicatives entre elles (multivarié)

Use Case de Machine Learning

Etape 3 : Preprocessing pour obtenir un dataframe prêt pour la Modélisation

Etape 4 : Modélisation

- définition du problème
- entraînement et comparaison de modèles baselines
- complexification des modèles
- optimisation des modèles (hyperparamètres, déséquilibre des classes)

Use Case de Machine Learning

Etape 5 : Interprétabilité

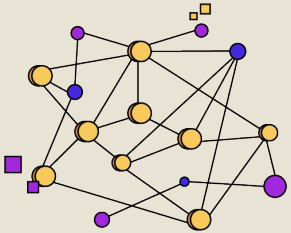
Etape 6 : Aller plus loin (MLOps)

- créer une pipeline pour le preprocessing et la Modélisation
- conteneuriser avec Docker pour assurer la reproductibilité et les dépendances correctes dans les environnements de développement
- automatiser et déployer le modèle conteneurisé avec Kubernetes
- créer une API pour rendre accessible le modèle avec FastAPI
- monitorer et contrôler l'API en production

Use Case de Machine Learning

3. Vulgariser les résultats d'un point de vue Business

- pour un nouveau client : possibilité de prédire si il va quitter la banque ou non avec 87% de fiabilité
- pour un nouveau client : possibilité de se focaliser sur sa perte et d'être sûr qu'il va quitter la banque avec 74% de fiabilité
- facteurs impactant la décision du nouveau client : principalement ses caractéristiques bancaires (nombre de produits souscrits, solde et présence) et son âge
- stratégie Business : se focaliser sur les clients susceptibles de partir et leur proposer des offres pour les retenir



Partie 3 : Questions techniques

Questions techniques

1. Qu'est-ce qui différencie l'apprentissage supervisé de l'apprentissage non-supervisé ? Pouvez-vous citer un modèle appartenant à chacune de ces catégories ?

Questions techniques

2. Qu'est-ce que la médiane d'une série statistique et en quoi diffère-t-elle de la moyenne ?

Questions techniques

3. Qu'est-ce que la p-value d'un test statistique et en quoi peut-elle être utile dans un projet de Data Science ?

Questions techniques

4. Si il y a une corrélation parfaite entre deux variables numériques d'un dataframe, combien vaut le coefficient de corrélation de Pearson ?

Questions techniques

5. On suppose que l'on fait de nouveau face à un dataset bancaire. Cette fois-ci, on veut prédire les chances qu'un prêt soit accordé à un client en fonction de ses caractéristiques. La variable cible prend 4 modalités : "aucune chance", "peu de chance", "fortes chances" et "quasiment sur".
Y a-t-il besoin d'encoder la variable cible ? Si oui, avec quel type d'encodage ?

Questions techniques

6. Dans le but de travailler sur de grosses databases, nous utilisons PySpark. A quoi sert le SparkContext ?

Questions techniques

7. En SQL, quelle méthode est utilisée pour regrouper les résultats d'une requête SELECT en fonction des valeurs d'une colonne ?

Questions techniques

8. A quoi sert la standardisation des variables numériques et quand doit-elle être effectuée ?

Questions techniques

9. Pour quel problème est utilisé le modèle Lasso et quelles sont ses principales caractéristiques ?

Questions techniques

10. L'accuracy est-elle une bonne métrique pour évaluer les performances d'un problème de classification déséquilibré ? Pourquoi ?

Questions techniques

11. Quel est le principal avantage du modèle XGBoost parmi les modèles de boosting existants ?
-> traitement des valeurs manquantes NA

Questions techniques

12. Qu'est-ce que la validation croisée k-fold (k-fold cross validation) ?

-> La cross validation (validation croisée) permet d'évaluer les performances d'un modèle. On découpe le dataset en plusieurs sous-ensembles, on entraîne le modèle sur certains d'entre eux puis on évalue ses performances de prédictions sur les sous-ensembles restants.
La cross validation la plus utilisée est la k-fold cross validation : on découpe le dataset en k sous-ensembles, on entraine sur k-1 et on évalue sur le sous-ensemble restant. On réitère k fois de telle sorte que le modèle est évalué sur chaque sous-ensemble.

Questions techniques

13. Comment définiriez-vous le Deep Learning ?
-> Sous-ensemble de l'IA, Machine Learning automatisé grâce à des réseaux de neurones

Questions techniques

14. Quel type de réseau de neurone est particulièrement utilisé pour traiter des images ? Quelle est son architecture classique ?
-> CNN avec couches de convolution, couches de pooling et couches denses

Questions techniques

15. A quoi sert la génération de tokens dans les modèles de NLP comme celui de ChatGPT ?
-> Découpe le texte en des éléments à preprocesser

Questions techniques

16. Les modèles ResNET sont particulièrement utilisés pour la classification d'images. Ils sont aussi très performants :

a) en traduction de langage
b) en prédiction de Séries Temporelles spécifiques
c) en analyse d'images médicales

-> c)

Questions techniques

17. Qu'est-ce que le dropout ?
-> méthode de réduction d'overfitting en Deep Learning pour les couches denses : le dropout désactive aléatoirement les neurones de la couche dense avec une certaine probabilité

Questions techniques

18. Quand et pourquoi utiliser des techniques de réduction de dimension ?
-> réduction de dimension pour les dataframes de très larges dimensions (plus de colonnes que de lignes ou nombre de colonnes dans l'ordre de la centaine/millier) afin de réduire l'overfitting, de faire de la DataVizualization et de diminuer le temps d'entrainement des modèles

Questions techniques

19. Votre dataset contient 30% de valeurs manquantes. Comment le traitez-vous ?
-> NA dans la variable cible : suppression des lignes
- > NA dans une variable explicative : calcul du taux de NA dans la variable pour aviser (suppression de la variable si taux très élevé, remplacement si taux modéré, suppression des lignes si taux faible)
-> prise de décision en fonction du nombre d'observations dans le dataframe

Questions techniques

20. Donnez un exemple de situation dans laquelle les faux négatifs sont tout aussi importants que les faux positifs.

-> contexte bancaire : prédire si un client va pouvoir rembourser un prêt

faux positif : il ne rembourse pas et la banque perd de l'argent sans avoir pris en compte le risque de solvabilité

faux négatif : il aurait pu rembourser et la banque perd de l'argent sur les potentiels intérêts