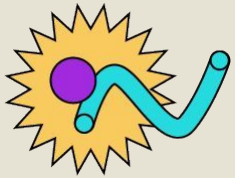
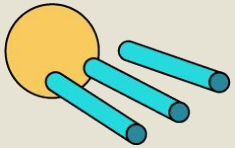
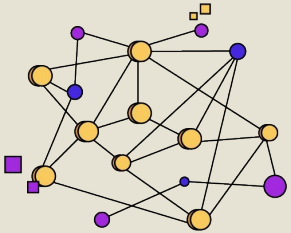


Technical Week Novembre 2024 : Test Technique Data Science



DataScientest



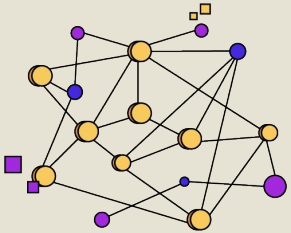
Déroulé d’un test technique DS

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. Écrivez une fonction qui prend en entrée une chaîne de caractères (un mot ou une phrase, supposée sans accents ni ponctuation) et qui renvoie **True si la chaîne est un palindrome, et False sinon**.

Rappel : Un palindrome est un mot ou une phrase qui peut se lire dans les deux sens. Par exemple, ‘kayak’ est un palindrome. De même, ‘La mariee ira mal’ est un palindrome.

2. Écrivez une fonction qui renvoie la **moyenne des chiffres composant un nombre**. Par exemple pour le nombre 1234, la fonction devra renvoyer 2,5.
3. Écrivez une fonction qui prend en entrée une liste et qui **renvoie l'ensemble des nombres premiers** présents dans cette liste.

Rappel : Un nombre premier est un nombre qui n’est divisible que par 1 et par lui-même. Notez que 1 n’est pas un nombre premier. De plus, un entier n ne peut pas être divisible par un nombre supérieur à $\sqrt{n}$.

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 `weight_change_dataset.csv'` sous forme de dataframe Pandas et visualisez les premières lignes.
3. Affichez les informations de l'ensemble des hommes ayant perdu du poids.
4. Affichez les informations de la 16ème personne la plus âgée.
5. Comptez le nombre de participants qui ont plus de 30 ans et une qualité de sommeil 'excellent' ou ‘good’
6. Comparez le poids médian des hommes et des femmes
7. Affichez la courbe de densité de la variable Age en différenciant selon le genre.
8. La difficulté à perdre du poids varie-t-elle avec l'âge, la qualité du sommeil et le niveau de stress d'après nos données ? Utilisez un graphique pour explorer ces relations.

Programmation sur Python

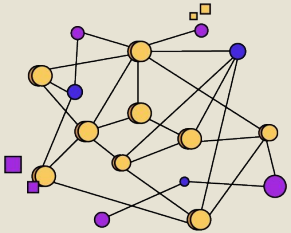
- savoir définir une fonction Python et notamment les fonctions usuelles (nombres premiers, puissance, suite de Fibonacci)

- 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 une étude de McKinsey, la satisfaction client est un facteur déterminant pour le succès des entreprises, en particulier dans le secteur de la vente au détail. Une satisfaction client élevée est souvent corrélée à la fidélité et aux recommandations des clients, tandis qu'une insatisfaction peut entraîner des pertes de clientèle et des avis négatifs en ligne, influençant ainsi l'image de marque.

Dans ce contexte, vous travaillez pour une chaîne de magasins de détail Nike, et votre base de données comprend des avis clients spécifiques aux produits et services Nike. L’objectif est d’analyser ces avis pour mieux comprendre les facteurs qui influencent la satisfaction client et pour identifier les points à améliorer afin de renforcer l'expérience d'achat et d'augmenter la fidélité des clients.

Question : En tant que Data Scientist, de quelles données auriez-vous besoin pour mener cette analyse de manière exhaustive ? Décrivez également votre processus de résolution pour identifier les leviers d’amélioration de la satisfaction client.

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 Métier : Quelle stratégie adopter pour améliorer la satisfaction client et renforcer la fidélité dans le contexte des magasins Nike ?

Problématique data : Comment segmenter les avis en thèmes principaux ? Comment prédire la satisfaction ou l'insatisfaction des clients en fonction de leurs messages ?

Approches :

- Clustering pour identifier les thèmes récurrents dans les avis clients.
- Classification pour prédire la satisfaction ou l'insatisfaction à partir du contenu des avis.

Données nécessaires : Avis clients sur les produits et services Nike incluant les notes, les commentaires, et des données sur les clients et les produits (si disponibles).

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

Variables disponibles :

- **Review** : Contenu de l'avis posté par le client
- **Rating** : Note attribuée par le client
- **Country** : Pays d'origine du client
- **user_avis** : Nombre d'avis publié par le client
- **Date** : Date à laquelle l'avis a été publié

A vérifier :

- Types des variables
- Valeurs manquantes
- Présence de duplicates
- Valeurs aberrantes

Use Case de Machine Learning

Etape 2 : Analyse de Données avec la DataVizualization

Objectif : Comprendre la répartition des avis clients et identifier les relations entre les variables pour guider l'analyse de satisfaction.

- **Distribution de la variable cible (univariée)** : Visualiser la distribution des notes de satisfaction (Rating) pour évaluer la répartition entre avis positifs et négatifs.
- **Analyse descriptive des variables (univariée)** : Étudier les caractéristiques des avis clients (pays d'origine, dates, etc.) pour mieux cerner le contexte des retours.
- **Analyse multivariée** : Explorer l'impact des différentes variables sur la note de satisfaction et analyser les corrélations potentielles entre ces variables explicatives.

Use Case de Machine Learning

Etape 3 : Preprocessing

Objectif : 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

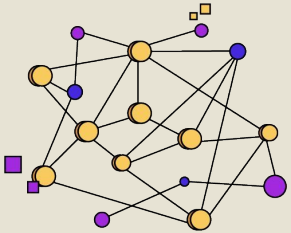
Etape 6 : Aller plus loin (MLOps)

- 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

- Prédiction de satisfaction : Pour un nouvel avis client, possibilité de prédire si le client est satisfait ou insatisfait avec un taux de fiabilité de 90%.
- Facteurs impactant la satisfaction client : Les facteurs influençant le plus la satisfaction sont principalement la qualité perçue des produits, le service client, et l'expérience d'achat en magasin ou en ligne.
- Stratégie Business : Se concentrer sur les clients potentiellement insatisfaits en leur proposant des solutions ou des offres personnalisées pour améliorer leur expérience et renforcer leur fidélité.



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 ?

Supervisé : variable cible à prédire ->
Régression/Classification (Random Forest par exemple)
Non-Supervisé : pas de variable cible à prédire -> Clustering (KMeans par exemple)

Questions techniques

2. Que peut-on dire de la moyenne et de la médiane dans le cas d'une distribution Normale ?

Elles sont égales

Questions techniques

3. Qu’est-ce que l’hypothèse nulle dans le cadre d’un test statistique ?

L’hypothèse à rejeter en fonction de la p-value

Questions techniques

4. Dans quels cas utilise-t-on le test du Chi-deux, et que permet-il de vérifier ?

Test de dépendance entre 2 variables catégorielles dans un but de feature selection

Questions techniques

5. Quelles hypothèses doivent être vérifiées concernant les erreurs d'une Régression Linéaire ?

Erreurs centrées, de même variance, indépendantes et de loi Normale

Questions techniques

6. Quelles méthodes permettent de transformer un modèle de classification binaire en un modèle de classification multiclasse ?

One VS One ou One VS All

Questions techniques

7. Quel est le rôle de la standardisation des variables numériques ? Quand est-il recommandé de l'appliquer ?

Mettre toutes les variables à la même échelle

Questions techniques

8. L'accuracy est-elle une métrique appropriée pour évaluer les performances dans un problème de classification déséquilibrée ?

Non car ne renvoie pas les résultats classe par classe
-> precision, recall, F1-score

Questions techniques

9. Pour quel type de problème utilise-t-on le modèle Lasso ? Quelles en sont les principales caractéristiques ?

Problème de régression :

- pénalité L1
- réduit l'overfitting
- permet la sélection de variables

Questions techniques

10. Qu’est-ce que la validation croisée stratifiée k-fold (stratified k-fold cross validation) ?

Technique d’évaluation des performances d’un modèle : k sous-ensembles -> (k-1) en entrainement et le dernier en test et on répète le processus k fois

Questions techniques

11. Quand et pourquoi utiliser des techniques de réduction de dimension ?

Pour réduire le temps d’entrainement et l’overfitting quand le dataframe a un trop grand nombre de variables en colonnes

Questions techniques

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

GROUP BY

Questions techniques

13. Comment définiriez-vous le Deep Learning ? Quelles en sont les principales branches ?

Sous-ensemble de l'IA basé sur des réseaux de neurones
-> MLP pour données tabulaires
-> Computer Vision pour données images
-> NLP pour données textuelles

Questions techniques

14. Comment fonctionne l'extraction de caractéristiques dans le cadre d'un réseau de neurones convolutif (CNN) ?

Couches de convolution avec kernel/filtre de convolution qui prend en compte les 3 dimensions de l’image et en extrait les caractéristiques

Questions techniques

15. Comment fonctionne le dropout et quel est son objectif ?

Permet de réduire l’overfitting en désactivant chaque neurone avec une certaine probabilité

https://survey.survicate.com/71b9f56833d456cb/?p=hubspot&first_name={{contact.firstname}}&last_name={{contact.lastname}}&email={{contact.email}}