Cdiscount

Test

Nom : …….

Prénom : ………

Date : ………….

**Merci de répondre aux questions suivantes en justifiant votre solution. Au besoin, vous expliciterez clairement vos hypothèses et vous les discuterez.**

**Préambule :**

Les questions ici posées n’attendent nullement une réponse stricte et définie. Elles sont avant tout posées pour voir votre niveau de maturité sur les différentes thématiques. Quelques phrases suffisent pour répondre à la plupart. Néanmoins, nous sommes bien conscients que la pluralité des sujets abordés rend difficile la réponse à toutes les questions.

Nous avons donc étiqueté quelques questions comme optionnelles : ces questions sont les plus avancées et si le candidat n’est pas à l’aise avec la thématique, il pourra l’abandonner.

**Probabilité/Statistiques**

***Question 1****: Quels seraient, selon vous, les facteurs de saisonnalité du Chiffre d’Affaires de cDiscount ?* (maximum 5 lignes)

***Question 2****: Comment désaisonnaliseriez-vous cette série? (maximum 10 lignes)*

***Question 3****: Mon taux de croissance de CA est de 15%. Comment pouvez-vous l’expliquer ? (maximum 10 lignes)*

***Question 4****: Toutes les semaines, il y a une probabilité de 10% d’avoir une baisse du CA. Quelle est la fréquence de l’évènement quatre semaines de baisse ?*

***Question 5****: j’achète sur une place de marché des visites. Le segment « hommes » achète avec une probabilité de 5%, le segment « femmes » avec une probabilité de 2%. Sachant que je paye l’acquisition de 100 visites hommes 50 euros, quel est le prix pour l’acquisition de 100 visites « femmes » ?*

***Question 6 (optionnelle)****: j’ai réussi à estimer un taux de transformation (nombre achats/nombre de visites) avec une précision de 1% à l’aide de 15 000 visites. De combien de visites ai-je besoin pour avoir une précision dix fois supérieure, à savoir 0,1% (1 pour mille) ?*

***Question 7******(optionnelle)****: Un tirage AB (probabilité d’être dans le groupe test est égale à 0.5) est réalisé pour un échantillon de taille 1 million. Le pourcentage observé d’individus dans la population test est de 48%. Est-ce normal ?*

***Question 8 (optionnelle)****: j’ai testé le taux de transformation sur deux versions du site lors d’un test AB (test randomisé). J’obtiens les résultats suivants :*

|  |  |  |
| --- | --- | --- |
|  | *Groupe Contrôle* | *Groupe Test* |
| *Taille Groupe* | *117415* | *117284* |
| *Taux de transformation observé* | *7,07%* | *9,36%* |

*Quel est l’intervalle de confiance du gain incrémental (Taux de transformation TEST – Taux de transformation CONTROLE) ?*

***Programmation théorique***

***Question 10****: Qu’est-ce que le polymorphisme ? Si vous l’avez-vous mis à contribution dans vos programmes, donnez un exemple.*

***Question 9*** ***(optionnelle)****: Détaillez les différences de gestion de la mémoire entre le C , le C++ et (Java,C#,Python). Argumentez sur les avantages et inconvénients de chacun.*

***Question 11 (optionnelle)****: Pouvez-vous citer des « Design Patterns » ? Si vous les avez mis à contribution dans vos programmes, donnez des exemples pour chacun d’entre eux.*

***Question 12******(optionnelle)****: Devant la diversité des algorithmes statistiques/machine learning, quelle est votre bibliothèque de prédilection pour prototyper vos solutions (en R ou Python ou Matlab) ?*

***Programmes demandés***

***Question 13****: calculez les coefficients de la régression linéaire de la variable y sur les variables x et la constante pour les individus dont la variable réponse vaut Y. Fichier regression.csv.Joindre le programme*

***Question 14****: le taux de transformation p2 est-il significativement supérieur au taux de transformation p1.Si oui, avec quelle fiabilité ? Fichier testmoyenne.csv.* Joindre le programme.

***Question 15******(optionnelle)****: sur la page* <http://stat-computing.org/dataexpo/2009/the-data.html>, *vous trouverez des données de transports aériens. Calculez les coefficients de la régression linéaire de la variable DepDelay sur les variables DayOfWeek, et heure de la semaine. Expliquez comment régresser des vecteurs de taille supérieurs à votre mémoire. Joindre le programme.*

***Question 16******(optionnelle)****: Construisez programmatiquement le graphe du site cdiscount.com, calculez le degré moyen sortant et entrant des pages (moyenne du nombre liens entrant la page, respectivement sortant) et calculez le page rank interne pour chaque page. Donnez les 10 pages les mieux rankées.* *Joindre le programme.*

***Question 17******(optionnelle)****: une méthode classique de recommandation est la technique dite de filtrage collaboratif. Il consiste à proposer à un client i les produits achetés par les clients ayant achetés. Les produits recommandés sont classés par fréquence d’apparition. Proposez un algorithme et donner sa complexité ? Votre algorithme est-il parallélisable ? Si non, comment faire ?*

***Question 18****: quelles pistes d’amélioration proposeriez-vous pour améliorer la personnalisation/recommandation sur cdiscount.com ?*

***Big Data***

***Question 19******(optionnelle)****: Pouvez-vous détailler les principales évolutions de la plateforme Hadoop. Cette évolution sera mise en perspective avec les différentes solutions s’étant développées autour pour l’exploiter.*

***Question 20 (optionnelle)****: Devant l’optique de monter en charge sur un volume de données « Big Data », comment imaginez-vous porter vos algorithmes de Machine Learning sur une architecture Big Data  (citer plusieurs frameworks de Machine Learning sur une architecture distribuée) ?*