# Stages Data Science et Data Analyse

# Embarquez à bord de la direction Data

Et venez partager votre passion avec une équipe pluridisciplinaire de plus de 30 data scientists et analysts en travaillant sur de nouveaux gisements de valorisation de la donnée.

# Les missions qui vous seront confiées

Rattachées à la direction Data Science, Innovation et Transformation, les équipes Data contribuent à optimiser les performances d’un nombre grandissant de services au sein de l’entreprise. Historiquement focalisés autour de la pertinence du moteur de recherche interne, les data scientists et data analysts abordent désormais des sujets tels que les systèmes de recommandation, l’acquisition de trafic ou encore le paiement en 4 fois. Pour cela, un large spectre de méthodes mathématiques sont appliquées aux centaines de téraoctets de données que l’infrastructure big data met à disposition, en forte interaction avec les services métiers qui pilotent l’activité et définissent les priorités business.

Vous travaillerez sur l’un des sujets ci-dessous, colorés Data Analyse (DA) ou Data Science (DS).

* Personnalisation du parcours de recouvrement (Data Analyse/Data Science)
* Amélioration de la détection de la fraude à la contestation de livraison (Data Science)
* Le texte comme vecteur de qualification des produits (Data Science)
* Mesure de la cohérence sémantique entre texte et image (Data Science)
* Amélioration du moteur de recherche interne (Data Science)
* Advanced Word Embedding of E-commerce queries (Data Science)
* Algorithme de recommandation de produits sponsorisés (Data Science)
* Amélioration de la résolution des images (Data Science)
* Matching de produits (Data Science)
* Pilotage du pricing de nos produits (Data Analyse/Science)

Vous trouverez ci-dessous une description plus détaillée des sujets proposés.

**Personnalisation du parcours de recouvrement (Data Analyse/Data Science)**

Ce stage est proposé par l'équipe scoring, responsable du score de solvabilité calculé en temps réel sur 120,000 pages panier par jour, qui détermine à quels clients Cdiscount propose le paiement à crédit (étalé en quatre mensualités). Certains clients tombent toutefois en défaut de paiement, et un processus de recouvrement de créance est alors lancé. Le parcours de recouvrement est identique pour tous, qu’il s’agisse d’un client fidèle qui a simplement changé de CB, ou d’un fraudeur multirécidiviste.

L'objectif serait donc de proposer une personnalisation du parcours de recouvrement afin d’en améliorer l’efficacité et l’expérience client.

Une piste possible serait d’utiliser les outils de score développés par l’équipe pour tenter d’identifier les clients les plus appétents à certains types de recouvrement.

Ainsi au cours de ce stage, vous appliquerez des méthodes avancées de Machine Learning sur des grandes volumétries constituées de données clients, de parcours de navigation et de transactions, et pourrez tester vos approches à grande échelle au moyen de tests AB.

**Amélioration de la détection de la fraude à la contestation de livraison (Data Science)**

La fraude à la contestation de livraison est en forte augmentation depuis le premier confinement début 2020 et la fin de la signature lors de la réception des colis. La grande partie de ces contestations sont légitimes et doivent donner lieu à une compensation, mais il peut également s’agir d’une technique d’arnaque très en vogue. Détecter le plus tôt possible ces fraudeurs est un enjeu important pour Cdiscount.

L’objectif sera donc de proposer un algorithme de détection de fraude dans un contexte de données mal labelisées. Il pourra être intéressant d’explorer les méthodes d’apprentissage semi supervisé, et on pourra employer par exemple des bases de données orientées graphe afin de construire un ensemble riche de features et tester plusieurs algorithmes de machine learning.

Ainsi au cours de ce stage, vous appliquerez des méthodes avancées de Machine Learning sur des grandes volumétries constituées de données clients, de parcours de navigation et de transactions, et pourrez tester vos approches à grande échelle au moyen de tests AB.

**Le texte comme vecteur de qualification des produits (Data Science)**

Depuis l’ouverture de la place de marché (Marketplace) sur le site de Cdiscount, le volume d’offres proposées par les vendeurs tiers sur notre site n’a cessé de croître. Aujourd’hui, ce sont plus de 100M d’offres qui sont accessibles en permanence par nos clients. Si ce foisonnement nous permet de répondre à tout type de besoin, il pousse aux limites notre capacité à garantir la qualité de notre catalogue.

L’équipe Qualiscore a pour mission de développer un indicateur permettant de contrôler la qualité des offres présentées à nos utilisateurs.

 Problématique : Comment estimer la qualité des fiches produit en se basant sur le texte ?

 Objectif : Trouver, par des méthodes de NLP, une façon d'attribuer aux fiches produits un score de qualité sémantique.

 Pistes : L'idée est de tester plusieurs approches Machine Learning de traitement du langage naturelle afin de déterminer la cohérence et la qualité sémantique des fiches produits. Quelques pistes de réflexions pourraient être :

Le texte du produit a-t-il un syntaxe correcte ?

Les différents éléments sémantiques d’une fiche produit sont-ils en accord ? (thématique, sens, valeurs) ?

Est-on capable de détecter la marque et le modèle d’un produit à partir de son libellé ?

Ces caractéristiques sémantiques ont-elles un impact sur la transformation des produits ?

Ces question conduisent à mener des travaux variés et prospectifs sur l’utilisation d’une ou d’un ensemble d'approche NLP (des modèles classiques aux modèles correspondant à l'état de l'art) et de les mettre en œuvre / optimiser dans le contexte très particulier du langage e-commerce.

L’aspect technique et opérationnel de la solution ne sera pas à négliger : est-on capable de mettre en œuvre les solutions proposées à l’échelle du catalogue Cdiscount.

**Mesure de la cohérence sémantique entre texte et image (Data Science)**

Depuis l’ouverture de la place de marché (Marketplace) sur le site de Cdiscount, le volume d’offres proposées par les vendeurs tiers sur notre site n’a cessé de croître. Aujourd’hui, ce sont plus de 100M d’offres qui sont accessibles en permanence par nos clients. Si ce foisonnement nous permet de répondre à tout type de besoin, il pousse aux limites notre capacité à garantir la qualité de notre catalogue.

L’équipe Qualiscore a pour mission de développer un indicateur permettant de contrôler la qualité des offres présentées à nos utilisateurs.

Problématique : Démontrer si l’incohérence sémantique entre texte et images d’une fiche produit (FP) représente un vrai irritant pour le client et trouver des méthodes pour quantifier celle-là de façon fiable.

Objectif : L’objectif de ce stage est d’étudier des approches possibles pour mesurer la cohérence sémantique existante entre le texte (en particulier le libellé) et les images d’une FP. Des questions auxquelles on pourrait essayer de répondre sont :

Est-ce que le fait d’observer une forte incohérence entre texte et images d’une FP représente un irritant pour le client ?

Laquelle, parmi les deux sources d’information, a généralement « raison » sur la nature du produit ? Est-elle la même source que le client privilégie ?

Y a-t-il des catégories (univers) particulièrement « sensibles » ?

L’aspect technique et opérationnel de la solution ne sera pas à négliger : est-on capable de mettre en œuvre les solutions proposées à l’échelle du catalogue Cdiscount.

## Amélioration du moteur de recherche interne (Data Science)

Chaque jour des millions de fois, notre moteur de recherche interne est interrogé et doit alors choisir, parmi un catalogue de 80 millions de produits, lesquels présenter à nos clients. Ce domaine de la recherche documentaire avec un fort retour implicite (ce que les utilisateurs regardent ou achètent in fine) est l'objet d'une attention académique constante depuis l'avènement d'internet. Parmi les avancées récentes, citons par exemple le modèle LambdaMART basé sur des forêts aléatoires boostées [1], ou RankNet basé sur des réseaux de neurones [2, 3], qui appartiennent à la famille des algorithmes "Learning to rank" (LTR).

Au cours de ce stage, on s'attachera ainsi à expérimenter les modèles LTR à l'état de l'art pour améliorer la pertinence de notre moteur de recherche. Les travaux porteront potentiellement sur toutes les étapes de construction des modèles. D'abord sur les données injectées en entrées (les "features") : on dispose ici d'un vaste réservoir à tester, entre le retour implicite, les performances commerciales passées, le texte des fiches produits ou encore les avis clients. Ensuite sur l'architecture générale et les paramètres et hyperparamètres des modèles, que l'on peut ajuster à l'infini. Enfin sur la sortie des modèles, qui est utilisée pour trier les produits par pertinence : une question est ici d'étudier si l'on peut recalibrer ce score de sortie de façon monotone pour le transformer en une quantité interprétable, comme l'espérance des ventes réalisées.

Si les résultats sont concluants, vous aurez l’opportunité de tester vos approches à grande échelle sur notre moteur de recherche interne, via un test AB.

[1] Burges, C.J.C. et al., 2008, Ranking, Boosting, and Model Adaptation

[2] Lazier, A. et al., 2005, Learning to Rank using Gradient Descent

[3] Köppel, M. et al., 2019, Pairwise Learning to Rank by Neural Networks Revisited: Reconstruction, Theoretical Analysis and Practical Performance

## Advanced Word Embedding of E-commerce queries (Data Science)

L’équipe data trafic travaille à améliorer la position de Cdiscount dans les pages de résultats gratuits et payants des moteurs de recherche externes, notamment Google et Bing. Elle manipule au quotidien des millions de requêtes utilisateurs et doit faire en sorte que le site Cdiscount puisse y répondre au mieux : a-t-on une page pertinente qui répond à la requête ? Cette page est-elle bien placée dans la structure du site ? Afin de pouvoir gérer automatiquement cette volumétrie, il est fondamental de pouvoir déterminer de manière fiable la proximité sémantique entre plusieurs requêtes ainsi qu’entre une requête et le contenu d’une page.

Les méthodes usuelles de type Bag Of Words, TF-IDF ou les word embedding classiques ne répondent que très partiellement à ce besoin. On expérimentera l’apport des architectures modernes introduites dans le domaine du NLP (BERT, GPT, …), que l’on entraînera et testera sur le vaste catalogue de données textuelles et de positionnement sur les moteurs de recherche externes dont nous disposons, afin de capturer au mieux les spécificités sémantiques liées notre activité de e-commerçant.

## Algorithme de recommandation de produits sponsorisés (Data Science)

L’équipe data en charge des produits sponsorisés a pour mission de permettre aux différents vendeurs et marques du site Cdiscount de bénéficier de davantage de visibilité pour leurs produits. Pour cela un certain nombre d’emplacements sponsorisés leur sont dédiés sur le site. L’équipe de data scientists a en charge de réaliser l’association entre un produit sponsorisé et un emplacement. Cette association se fait sur des critères de pertinence du produit à l’emplacement mais également des enchères (au clic) et des budgets des différents annonceurs.

Afin d’optimiser toujours plus loin l’adéquation entre l’offre disponible et la demande de nos utilisateurs, nous souhaitons développer des algorithmes de personnalisation en temps réel afin d’identifier les produits susceptibles d’intéresser le plus nos utilisateurs tout en générant un revenu publicitaire associé. Pour cela nous serons amenés à utiliser l’historique de navigation et d’achat du visiteur ou client afin de proposer les produits les plus pertinents.

L’objectif de ce stage est dans un premier temps de faire une revue non exhaustive des techniques récentes et performantes de recommandation produit, tout en recoupant avec les techniques pré existantes côté Cdiscount. Dans un deuxième temps, le candidat aura l’opportunité de mettre en œuvre son algorithme et de faire évoluer sa solution sur les données du site.

**Amélioration de la résolution des images (Data Science)**

Ce stage est proposé par l’équipe data science consulting, responsable de la réalisation de projet courts pour les différentes transversales de Cdiscount, dans le cadre d’un projet existant avec les équipes du front.

Avec l’ouverture du site à des vendeurs tiers et l’accroissement continu du catalogue de produit, un nouvel enjeu est de garantir les standards de qualité des produits de Cdiscount pour toutes les fiches produit présentes sur le site. Un des éléments importants en sont les images, qui parfois sont malheureusement de mauvaise qualité.

L’objectif de ce stage est l’amélioration des techniques en place de super-résolution pour retravailler au maximum les images du site en centrant et agrandissant les images des fiches les plus consultées.

La première partie du stage, de deux mois environ, sera réservée à une analyse bibliographique et un benchmark des différentes techniques présentes dans la littérature. La suite sera la prise en main de l’algorithme existant et son amélioration avec une validation de la part des équipes responsables du front pour une mise en production avant la fin du stage.

**Matching de produits (Data Science)**

Ce stage est proposé par l’équipe data science consulting, responsable de la réalisation de projet courts pour les différentes transversales de Cdiscount, dans le cadre d’un projet existant avec les équipes de l’offre.

Lorsqu’un nouveau produit est présenté par un fournisseur ou un tiers, il n’est pas aisé de savoir s’il est déjà présent dans le catalogue volumineux de Cdiscount et des vendeurs tiers affiliés à la Marketplace. Et c’est pour répondre à ce besoin que l’équipe a mis en place un algorithme de matching de produits utilisant une approche sémantique avec un BERT.

Nous souhaitons dans le cadre de ce stage pousser le matching visuel, éventuellement sur certaines typologies de produits, pour améliorer notre moteur sur les cas où le titre est peu informatif (par exemple « chaussure marron » ou « adidas superstar taille 38 »).

La première partie du stage, de deux mois environ, sera réservée à une analyse bibliographique et un benchmark des différentes techniques présentes dans la littérature. La suite sera la prise en main de l’algorithme existant et son amélioration avec une validation de la part des équipes de l’offre pour une mise en production avant la fin du stage.

## Pilotage du pricing de nos produits (Data Analyse/Science)

Trouver le meilleur prix possible pour le consommateur et pour Cdiscount est un enjeu central de notre plateforme de e-Commerce afin d’établir une relation et une confiance sur le long terme avec nos clients, nos fournisseurs et en améliorant notre rentabilité. C’est cependant un exercice notoirement complexe, tant les réactions des ventes sont bruitées par des facteurs autres que le prix : météo, trafic sur le site, variations du prix des concurrents, visibilité des produits sur notre site etc. La réalité est souvent bien loin de la simple formule de l’élasticité-prix de la demande.

Cdiscount déploie des efforts importants pour améliorer le pilotage du pricing. Dans ce cadre, le travail du candidat pourra porter sur plusieurs axes parmi les suivants, en fonction de ses appétences, de sa spécialité (data science ou analyse) et des priorités de l’entreprise au démarrage du stage.

D’abord, de nombreux aspects de la modélisation de l’impact d’un mouvement de prix sur les ventes restent à affiner : comment estimer des élasticités fiables, et sur quels agrégats de produits ? A quelle impulsion des ventes faut-il s’attendre lorsqu’on se place sous le prix minimum du marché ? Comment modéliser les reports de ventes lors d’une hausse de prix ?

Ensuite, se pose la question de la stratégie de pilotage des prix. Avec un modèle des ventes fiable, le domaine de la recherche opérationnelle offre un cadre naturel pour construire une stratégie qui intègre les contraintes de l’entreprise. D’autres pistes exploratoires sont envisageables, comme l’apprentissage par renforcement pour expérimenter des variations de prix dans un environnement changeant.

Enfin, viennent le suivi et l’analyse des actions. Comme les ventes réagissent à de nombreux facteurs, il est important de proposer une méthodologie de suivi et des métriques capables de démêler les effets pour isoler l’impact des mouvements de prix.

Si les résultats expérimentaux sont encourageants, le candidat pourra tester sa méthodologie à grande échelle en conditions réelles.

# Profil recherché

De formation supérieure de type Bac+4/5 ou équivalent en école d’ingénieur spécialisé en data science. Vous manifestez un intérêt pour le e-commerce et le web et vous êtes d’un naturel rigoureux, motivé et curieux.

Vous êtes à l’aise avec Python, SQL et avez potentiellement des notions en big data. Vous vous épanouissez dans les sujets analytiques et leur application à des cas concrets de l’entreprise. Vous avez une appétence certaine pour les sujets techniques et les problèmes complexes. Tout cela vous permet de transformer rapidement des idées théoriques en algorithme fonctionnel.

Vous cherchez une société où vous pourrez mettre en œuvre votre créativité et votre expertise pour relever nos défis quotidiens.

Vous vous reconnaissez ? Ne cherchez plus, postulez et partez à la conquête du web !

# Informations complémentaires

Type de contrat : Stage

Localisation : Bordeaux (33)

recrutement@cdiscount.com

# Comment postuler

Vous souhaitez encore plus d’informations ? N’hésitez pas à visiter notre espace recrutement : <https://emploi.cdiscount.com/>