STAGE – DATA SCIENTIST H/F

Embarquez à bord du Marketing:

Et venez booster notre trafic et nous aider à connaitre, attirer et fidéliser nos 16 millions de clients,

Les missions qui vous seront confiées :

Rattachée au service Marketing, la data science occupe une place grandissante chez Cdiscount. Des dizaines d'algorithmes sont actuellement en production sur le site, qui impactent l'expérience de millions d'utilisateurs sur des sujets aussi variés que la pertinence du moteur de recherche, les systèmes de recommandations, l'acquisition de trafic ou la qualité du catalogue de produits. Ces algorithmes se nourrissent de données de navigations, de transactions et de produits, stockées sur une plateforme big data Hadoop qui pèse plusieurs centaines de téraoctets. Aux côtés de data scientists seniors, ce stage vous permettra de travailler dans un environnement big data mature, en prenant conscience des problématiques du e-commerce.

Les sujets de stage que nous proposons couvrent bon nombre de domaines dans lesquels les data scientists travaillent aujourd'hui :

- **Recommandations par graphe** – Construction d'un graphe de produits à partir des caractéristiques de produits, afin de proposer des recommandations de montée en gamme.
- **Personnalisation des recommandations** – Optimisation des recommandations de produits par segment de client.
- **Scoring de fiches produits** – Construction et optimisation d'un score de qualité intrinsèque aux fiches produits.
- **Prévision des caractéristiques de produits** à partir du texte et des images, par apprentissage supervisé
- **Prévision des pages crawlées par Google** à partir des pages visitées par le robot Google durant les dernières heures, par apprentissage supervisé.
- **Optimisation du maillage –** Mise en place d'un algorithme mathématique de construction de graphe optimal.
- **Enchères par audiences** – Optimisation des enchères sur Google AdWords par segment de clients.
- **Prévision du score de qualité Google** – Prévision de l'impact du type de page et du texte de l'annonce sur le score de qualité calculé par Google AdWords.
- **Prévision des ventes** – Etude des données et des méthodes d'analyse de séries temporelles pour améliorer la prévision des ventes à venir, dans une optique d'optimisation logistique.
- **Prévision de la probabilité d'achat** à partir du parcours de navigation, par apprentissage supervisé.

Différentes autres tâches pourront être proposées, plus opérationnelles, permettant d'avoir une vision plus globale de l'activité des data scientists.

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.

Python, SQL et Linux n'ont plus de secret pour vous, vous vous épanouissez dans les sujets analytiques et la gestion de projet. 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.

Leader dans l'âme, 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/

Cdiscount · Recrute

Descriptions détaillées des sujets :

Recommandations par graphe :

L'objectif de ce stage consiste à élaborer un système de recommandation par graphe basé sur les propriétés des produits : mémoire disponible, taille d'écran, couleur, résolution etc. Vous chercherez, à partir d'un produit, à proposer des produits pertinents avec des caractéristiques différentes afin d'aider nos clients dans leur choix d'achats : par exemple à partir d'un smartphone proposer des téléphones de la même gamme avec moins de stockage, une meilleure résolution, plus légers… Votre algorithme devra donc être capable de justifier ses choix en comparant les caractéristiques des différents produits et ce de manière automatique. La solution fournie devra avoir une approche graphe : chaque produit est un nœud, à vous de trouver les liens qui les unissent.

Pour entraîner vos modèles, vous aurez accès à notre catalogue de 30 millions de produits avec toutes leurs propriétés associées et à un environnement technique robuste (plate-forme big data, serveurs de calcul dédiés, cloud computing). Vous pourrez éventuellement vous appuyer sur des données comportementales comme l'historique des produits consultés par session, qui permettent elles aussi de dériver une notion de graphe guidé par les utilisateurs.

Par un test A/B, vous pourrez mettre en production votre travail sur le système de recommandations du site internet, afin de tester sur des millions d'internautes l'appétence des produits proposés par votre algorithme.

Personnalisation des recommandations :

Dans une optique d'amélioration de l'expérience client, nous souhaiterions développer un système de recommandations destinées à répondre aux mieux aux attentes et besoins des internautes. Votre objectif sera donc de mettre au point un algorithme de clustering afin de les regrouper suivant leur navigation et d'autres critères que vous définirez.

Ensuite, pour personnaliser la recommandation des produits proposés sur le site, une deuxième étape visera à définir les produits les plus appropriés à chacun des clusters et de les présenter aux clients.

Par un test A/B, vous pourrez mettre en production votre travail sur le système de recommandations du site internet, afin de tester sur des millions d'internautes l'appétence des produits proposés par votre algorithme.

Pour ce faire vous serez intégré au sein de l'équipe Pertinence en charge du moteur de recherche et des recommandations produits sur le site et vous aurez accès à notre catalogue de 30 millions de produits avec toutes leurs propriétés associées et à un environnement technique robuste (plate-forme big data, serveurs de calcul dédiés, cloud computing).

Scoring de fiches produits :

Dans l'équipe data control, vous serez amené à travailler sur l'amélioration du scoring qualité des fiches produits. Ce score impacte directement les résultats du moteur; en effet un bon score donnera à un produit une meilleure visibilité dans les listes de recherche.

Ce stage s'articule autour de trois axes : tout d'abord, vous devrez définir différents marqueurs, prenant en compte les avis clients, la qualité de la description, ou encore des données techniques. Dans un second temps, vous construirez un score agrégé à partir des marqueurs individuels (en utilisant par exemple une approche bayésienne). Enfin, vous analyserez l'impact de ces différents marqueurs sur la conversion.

Prévision des caractéristiques de produits :

Une grande majorité des produits visibles sur notre site sont proposés par des vendeurs indépendants et les caractéristiques (couleur, matière…) ne sont pas toujours renseignées ce qui dégrade l'expérience utilisateur. En plus de ne pas être affichée lors de l'utilisation des filtres propres à ces attributs, une fiche produit incomplète peut être frustrante pour le client et dégrader sévèrement la probabilité de conversion de sa visite.

Ce stage s'articule autour de deux axes : tout d'abord, vous serez amené à analyser les relations entre le remplissage et les différents indicateurs (taux de transformation, vues sur la fiche produit…) pour guider le choix des attributs prépondérants pour chacune des catégories. Dans un second temps, vous devrez identifier et tester des algorithmes de traitement d'image (réseaux Deep) et de texte (Fine-grained Entity Type Classification) permettant de remplir ces attributs. Dans une démarche d'industrialisation de vos résultats, vous aurez l'occasion de les tester et d'en mesurer l'impact réel à l'aide d'un test A/B.

Prévision des pages crawlées par Google :

Dans ce stage nous nous intéressons à l'amélioration du crawl de Google sur le site Cdiscount. Le crawl est la première étape du processus d'indexation d'un moteur de recherche. Il permet de récupérer de l'information qui est ensuite organisée par les algorithmes des moteurs de recherches pour présenter les résultats les plus pertinents aux utilisateurs [1]. La compréhension, la maitrise et l'amélioration du processus de crawl est une problématique très importante pour l'équipe SEO et est un des facteurs impactant la visibilité de Cdiscount dans les moteurs de recherches.
Dans un premier temps nous souhaitons être capables de prédire l'ensemble des pages que le crawler de Google va explorer sur Cdiscount dans la prochaine heure. Le point de départ sera d'utiliser l'historique des logs serveurs pour détecter des patterns dans l'exploration du bot. Pour affiner le modèle vous aurez à disposition l'ensemble des données de Cdiscount, comme par exemple le graphe du site, le parcours des utilisateurs etc.
Dans un second temps nous souhaitons pouvoir utiliser ce modèle prédictif afin de piloter le crawl de Google et mettre en avant certaines pages du site. Le sujet pourra alors s'orienter sur des problématiques de type graphe afin d'adapter dynamiquement le maillage du site sur l'ensemble des futures pages crawlées.
Références :
[1] https://www.google.com/search/howsearchworks/crawling-indexing/

Optimisation du maillage :

Dans l'équipe référencement naturel, un des problèmes posés concerne l'optimisation du graphe constitué par les pages du site Cdiscount (nœuds) et les liens d'une page vers une autre (arêtes orientées). On sait que la qualité du maillage d'un site internet influence fortement le classement sur Google : il faut éviter les pages orphelines (sans lien entrant), minimiser la profondeur par rapport à la page d'accueil, éviter les liens entre des pages sans rapport…
Le but de ce stage sera justement d'identifier et implémenter les méthodes mathématiques capables de construire un graphe en optimisant explicitement une fonction objective de qualité. On part d'un ensemble de nœuds (pages) et de similarités entre les paires de nœuds, que l'on sait en pratique facilement calculer en se basant sur le texte, les produits présentés ou les sessions de navigation. La difficulté réside en ce qu'il faut poser des arêtes entre des nœuds similaires (métrique locale) tout en optimisant une métrique globale de graphe (par exemple, rayon du graphe ou excentricité par rapport à la page d'accueil), ce sur des millions de nœuds. Vous aurez l'occasion de tester votre algorithme sur une zone de maillage et un magasin donné du site de Cdiscount, afin d'observer l'effet sur le classement Google.

Enchères par audiences :

Google propose un service d'advertising (Google Ads), qui permet d'acheter un espace commercial (Ad) dans des espaces d'une page de résultats de recherche (typiquement en haut et en bas). La présence et le positionnement dans la page dépendent d'un mécanisme d'enchères qui met en concurrence plusieurs annonceurs sur chaque phrase-clé.
Afin d'optimiser le retour de l'investissement publicitaire, la connaissance du comportement client permet d'identifier des segments qui, pour une catégorie de produit donnée ou à la suite d'une certaine recherche Google, sont plus susceptibles de convertir, et sur lesquels on va placer des enchères plus élevées.
Au sein de l'équipe Acquisition, votre mission sera d'identifier une segmentation client automatique et dynamique (mise à jour régulièrement) qui permet de décrire le parc client actuel et d'anticiper le comportement d'un nouveau client, avec l'objectif d'améliorer les performances en termes de chiffre d'affaire et recrutement de nouveaux clients. Vous aurez l'occasion de tester votre segmentation et d'en mesurer l'impact réel à l'aide d'un test A/B.

Prévision du score de qualité Google :

Google propose un service d'advertising (Google Ads), qui permet d'acheter un espace commercial (Ad) dans des espaces d'une page de résultats de recherche (typiquement en haut et en bas). La présence et le positionnement dans la page dépendent d'un mécanisme d'enchères qui met en concurrence plusieurs annonceurs sur chaque phrase-clé.
Pour assurer des annonces pertinentes et utiles à l'utilisateur, Google affecte un score de qualité à chaque Ad soumise : ce score est une mesure globale qui dépend de plusieurs facteurs, comme la pertinence de l'annonce textuelle et de la page de redirection, et d'autres inhérents à la phrase-clé. Un score élevé permet d'améliorer son positionnement avec une enchère moindre.
Au sein de l'équipe Acquisition, votre mission sera d'identifier les axes d'amélioration de notre score de qualité sur un parc important de phrases-clés. Vous chercherez notamment à quantifier l'effet du texte de l'annonce et du type de page de redirection sur le score, par une approche qui pourra faire appel à des méthodes d'apprentissage supervisé. Le but sera donc de choisir la page de redirection et l'annonce optimales, avec pour objectif l'amélioration des performances en termes de trafic et chiffre d'affaire. Vous aurez la possibilité de tester votre approche et d'en mesurer l'impact au moyen d'un test A/B.

Prévision des ventes :

Les entreprises de commerce, qu'elles soient physiques ou numériques, réalisent des projections de leurs ventes pour optimiser la gestion des stocks : il s'agit d'éviter les ruptures, qui représentent du manque à gagner, aussi bien que l'excès de stocks, afin de minimiser les coûts d'entreposage.

Ce stage a pour but d'explorer les données et méthodes d'analyse des séries temporelles afin d'améliorer la prédiction des ventes des produits proposés par Cdiscount. Une première étape sera d'extraire des variables potentiellement explicatives de la quantité de données à disposition – ventes passées, données de navigation ou encore météo – et de quantifier leurs liens avec les ventes à venir à l'aide de méthodes d'apprentissage supervisé (modèles autorégressifs, forêts aléatoires…).

Il sera aussi important de se pencher sur des modèles plus spécifiques, adaptés à des cas particuliers comme le début ou la fin de vie d'un produit, les périodes de soldes ; de même que sur la prise en compte de cyclicités intrinsèques au e-commerce, liées aux week-ends, à la réception des salaires, aux vacances etc. Ce stage pourra éventuellement déboucher sur une thèse de type CIFRE.

Prévision de la probabilité d'achat :

La probabilité d'achat d'un client est un indicateur business clé pour les entreprises e-commerce. Si on peut facilement estimer cette probabilité de façon globale a posteriori, il est assez complexe de l'anticiper pour chaque client. Le but de ce stage est de mettre au point des méthodes visant à prédire la probabilité qu'un client achète en se basant sur les nombreuses données de navigations du site (plus de 2 millions de visites par jour !).

Une première phase sera l'extraction de variables pertinentes basées notamment sur le comportement utilisateur : quelle est la séquence de pages vues, quels types de produits sont consultés, quelle est la fréquence de connexion du client… On pourra également utiliser des variables « froides », comme l'historique de visite, d'achat, les interactions sur les emails, les publicités de Cdiscount etc. Ensuite, des méthodes d'apprentissage automatisé supervisé seront mises en œuvre afin d'utiliser ces variables pour prédire la probabilité d'achat du client. On pourra aussi tester des méthodes bayésiennes pour affiner notre prédiction au fil de la session de navigation.

Les applications potentielles de ces travaux sont nombreuses : éditions ciblées de bons de réduction, recommandations personnalisées… Vous aurez l'occasion de tester votre algorithme sur un cas d'usage concret au moyen d'un test A/B.

Cdiscount · RH