Fondamentaux de la Data Science

DESCRIPTION
Surfant sur la vague du Big Data, le data scientist joue un rôle clé dans la valorisation de données. Au-delà des paillettes, quel est son rôle, ses outils, sa méthodologie, ses "tips and tricks" ? Venez le découvrir au travers de cette initiation à la Data Science délivrée par des data scientists renommés qui vous apporteront l'expérience des compétitions de Data Science et leurs riches retours d'expérience des modèles réels qu'ils mettent en place chez leurs clients.

OBJECTIFS PEDAGOGIQUES

- Découvrir le monde de la Data Science et les grandes familles de problèmes
- Savoir modéliser un problème de Data Science
- Créer ses premières variables
- Constituer sa boite à outils de data scientist

PUBLIC CIBLE

- Analyste
- Statisticien
- Architecte
- Développeur

PRE-REQUIS
Connaissances de base en programmation ou scripting. Quelques souvenirs de statistiques sont un plus.

METHODE PEDAGOGIQUE
Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.

Stage pratique
Data Science

Code :
DSFDX

Durée :
3 jour(s) (21,00 heures)

Exposés : **60 %**
Cas pratiques : **30 %**
Echanges d'expérience : **10 %**

Inter-entreprises :
Prochaines sessions disponibles sur notre site web.
Tarif : 2 450,00 € HT / participant

Intra-entreprise :
Tarifs et dates sur demande.

PROFIL DES INTERVENANTS
Toutes nos formations sont animées par des consultants-formateurs expérimentés et reconnus par leurs pairs.

MODALITÉS D'ÉVALUATION ET FORMALISATION À L'ISSUE DE LA FORMATION
L'évaluation des acquis se fait tout au long de la session au travers des ateliers et des mises en pratique. Une évaluation à chaud sur la satisfaction des stagiaires est réalisée systématiquement en fin de session et une attestation de formation est délivrée aux participants mentionnant les objectifs de la formation, la nature, le programme et la durée de l'action de formation ainsi que la formalisation des acquis.

PROGRAMME PEDAGOGIQUE DETAILLE

Jour 1

INTRODUCTION AU BIG DATA

Qu'est-ce-que le Big Data ?

L'écosystème technologique du Big Data

INTRODUCTION À LA DATA SCIENCE

Le vocabulaire d'un problème de Data Science

De l'analyse statistique au machine learning

Overview des possibilités du machine learning

MODÉLISATION D'UN PROBLÈME

Input / ouput d'un problème de machine learning

Mise en pratique "OCR"

- Nous verrons comment modéliser le problème de la reconnaissance optique de caractère

IDENTIFIER LES FAMILLES D'ALGORITHMES DE MACHINE LEARNING

Analyse supervisée

Analyse non supervisée

Classification / régression

SOUS LE CAPOT DES ALGORITHMES : LA RÉGRESSION LINÉAIRE

Quelques rappels : fonction hypothèse, fonction convexe, optimisation

La construction de la fonction de coût

Méthode de minimisation : la descente de gradient

SOUS LE CAPOT DES ALGORITHMES : LA RÉGRESSION LOGISTIQUE

Frontière de décision

La construction d'une fonction de coût convexe pour la classification

LA BOITE À OUTIL DU DATA SCIENTIST

Introduction aux outils

Introduction à python, pandas et scikit-learn

CAS PRATIQUE N°1 : "PRÉDIRE LES SURVIVANTS DU TITANIC"

Exposé du problème

Première manipulation en python

Jour 2

RAPPELS ET RÉVISION DU JOUR 1

QU'EST-CE QU'UN BON MODÈLE ?

Cross-validation

Les métriques d'évaluation : precision, recall, ROC, MAPE, etc

LES PIÈGES DU MACHINE LEARNING

Overfitting ou sur-apprentissage

Biais vs variance

La régularisation : régression Ridge et Lasso

DATA CLEANING

Les types de données : catégorielles, continues, ordonnées, temporelles

Détection des outliers statistiques, des valeurs aberrantes

Stratégie pour les valeurs manquantes

Mise en pratique : "Remplissage des valeurs manquantes"

FEATURE ENGINEERING

Stratégies pour les variables non continues

Détecter et créer des variables discriminantes

CAS PRATIQUE N°2 : "PRÉDIRE LES SURVIVANTS DU TITANIC"

Identification et création des bonnes variables

Réalisation d'un premier modèle

Soumission sur Kaggle

DATA VISUALISATION

La visualisation pour comprendre les données : histogramme, scatter plot, etc

La visualisation pour comprendre les algorithmes : train / test loss, feature importance, etc

INTRODUCTION AUX MÉTHODES ENSEMBLISTES

Le modèle de base : l'arbre de décision, ses avantages et ses limites

Présentation des différentes stratégies ensemblistes : bagging, boosting,

OCTO Academy est une marque d'OCTO Technology, référencé au DATADOCK – 34, avenue de l'Opéra, 75002 Paris France - SAS au Capital de 509 525.30 €
Tél : +33 1 70 36 18 67 – Fax : +33 1 58 56 10 01 – academy@octo.com – https://www.octo.academy – academy@octo.com
SIRET : 418 166 096 00069 – RCS Paris B 418 166 096 - NAF 6202A – Organisme de formation N° 11 75 48 936 75

etc

Mise en pratique : "Retour sur le titanic"

- Utilisation d'une méthode ensembliste sur la base du précédent modèle

APPRENTISSAGE SEMI-SUPERVISÉ

Les grandes classes d'algorithmes non supervisées : clustering, PCA, etc

Mise en pratique : "Détection d'anomalies dans les prises de paris"

- Nous verrons comment un algorithme non supervisé permet de détecter des fraudes dans les prises de paris

Jour 3

RAPPELS ET RÉVISIONS

Synthèse des points abordés en journées 1 et 2

Approfondissement des sujets sélectionnés avec l'intervenant

MISE EN PRATIQUE

Le dernier jour est entièrement consacré à des mises en pratique

SÉLECTION ET PARTICIPATION À UNE COMPÉTITION

Le formateur sélectionnera une compétition en cours sur Kaggle ou datasciencenet qui sera démarrée en jour 3 par l'ensemble des participants

OCTO Academy est une marque d'OCTO Technology, référencé au DATADOCK – 34, avenue de l'Opéra, 75002 Paris France - SAS au Capital de 509 525.30 €
Tél : +33 1 70 36 18 67 – Fax : +33 1 58 56 10 01 – academy@octo.com – https://www.octo.academy – academy@octo.com
SIRET : 418 166 096 00069 – RCS Paris B 418 166 096 - NAF 6202A – Organisme de formation N° 11 75 48 936 75