Formation Data Scientist : Les fondamentaux de la Data Science

Objectifs de la formation Data Scientist :

Le métier de Data Scientist est apparu ces dernières années pour faire face à la multiplication des données, à la diversité de leurs formes et de leurs sources : le Big Data. Le rôle du Data Scientist : rendre les données exploitables, les traiter pour leur donner du sens et ainsi permettre à la direction générale d'adapter la stratégie de l'entreprise. C'est donc un acteur-clé aux compétences multiples. Cette formation *Data Scientist : Les fondamentaux de la Data Science* se concentre sur l'aspect technique de ce métier. Vous découvrirez les méthodes et les outils du Data Scientist et partagerez les retours d'expériences des formateurs. Des exercices pratiques et la participation à une compétition vous permettront d'expérimenter vos nouvelles connaissances.

Plus concrètement, à l'issue de cette formation vous serez aptes à :

- Découvrir le métier de Data Scientist et les grandes familles de problèmes
- Savoir modéliser un problème de Data Science
- Créer vos premières variables
- Constituer votre boîte à outils de Data Scientist
- Participer à une première compétition.

À qui s'adresse cette formation ?

Public :

Ce stage s'adresse aux Analystes, Statisticiens, Architectes, Développeurs.

Prérequis :

Pour suivre ce cours dans les meilleures conditions possibles, il vous faut avoir certaines connaissances de base en programmation ou Scripting, ainsi que quelques souvenirs de statistiques qui peuvent être un plus.

Contenu du cours Data Scientist:

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 métier de Data Scientist

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

Travaux Pratiques « OCR» :

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 boîte 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 révisions 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

Over fitting 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

Travaux Pratiques :

« 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, etc

Travaux Pratiques "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

Travaux Pratiques « Détection d'anomalies dans les prises de paris» :

Comment un algorithme non supervisé permet-il 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 datascience.net qui sera démarrée en jour 3 par l'ensemble des participants