

Formation Machine Learning et Data Science – Exploiter des modèles en production

Code : NTE07**Durée :** 3 jours**Classe :** Présentiel / à distance

Public

- Développeurs
- Data Scientists
- Ingénieurs Machine Learning
- Data Engineers / Data Ops

Prérequis

- Notions en Python
- Bases en statistiques
- Notions en Machine Learning
- Compréhension des systèmes Linux

Objectifs

- Déployer des modèles de Machine Learning en production
- Mettre en place le monitoring des modèles
- Gérer le cycle de vie des modèles (retrain, update)
- Mettre en œuvre les meilleures pratiques de MLOps et CI/CD

Programme détaillé

1- Introduction et rappels

- Les objectifs du machine learning : résoudre un problème
- Rappels de l'écosystème Data Science : langages, outils et pratiques
- Les raisons des échecs des passages en production
- Les principales difficultés rencontrées pour la mise en production de modèles

2- État de l'art de l'organisation des projets de Data Science

- Qui déploie les modèles et comment ?
- Les différents rôles : le data scientist, le data engineer, le data ops
- Les limites à cette organisation sur la réussite des projets data

3- État de l'art des solutions logicielles de déploiement de modèles de Machine Learning

- État de l'art : des déploiements sans outillage
- Limites techniques et coûts élevés
- Maintenabilité complexe
- Un exemple avec Tensorflow, PyTorch, scikit-learn, Python, Java
- Un exemple dans le Cloud (AWS, GCP, Azure)

4- Les bonnes pratiques liées aux métiers afin de réduire les risques d'échec

- Des critères d'acceptabilité (user acceptance) obligatoires par le métier : le premier pré-requis
- Des moyens de tester les algorithmes avec les métiers tous les deux jours
- Une communication permanente entre data scientist, data ops, IT, métiers...
- Identifier les données requises et leurs disponibilités

5- Des solutions techniques pour faciliter et améliorer les déploiements

- De nouveaux outils et logiciels : TFX, Mlflow, Kubeflow, Cloudera Data Science Workbench, Dataiku
- De nouvelles compétences : l'ingénieur machine learning, le data ops
- L'importance du choix d'un framework sur la continuité, la maintenabilité et l'utilisation d'un modèle

6- Nouveaux métiers et profils : de nouvelles compétences

- L'ingénieur machine learning
- Le data ops



Formation Machine Learning et Data Science – Exploiter des modèles en production

Code : NTE07

Durée : 3 jours

Classe : Présentiel / à distance

7- Le choix des outils

- État de l'art des outils / frameworks utilisés par les grandes sociétés (CAC40, grandes startups, GAFAM)
- Comment choisir un framework de Machine Learning ou de Data Science
- Impacts sur les coûts des projets
- Estimer le coût des outils (in)existants sur le projet
- Estimer le coût des outils de collaboration, de déploiement, de monitoring, etc. sur les projets data

8- Les critères d'un bon modèle pour la production

- Portabilité du modèle
- Scalabilité
- Utilisabilité par les applications métiers
- Testabilité

9- Les différents formats des modèles

- Pickle
- ONNX
- PMML
- POJO & MOJO

10- Déployer en production

- Entraîner le modèle en production : batch training, real time training
- Batch vs real-time prediction : impacts sur les performances, les évaluations, les infrastructures et les coûts
- Monitorer et mesurer les écarts en production
- Entraîner de façon continue les modèles (retrain, re-score, re-validate)
- Mesurer le drift des modèles (automated model checking)
- Remplacer l'ancien modèle en production par le nouveau modèle amélioré

11- Mesurer la réussite d'un déploiement en production

- Comment mesurer la réussite d'un déploiement ?
- La mesure de la réussite des déploiements en CI sur un projet data
- Les métriques

