Quantmetry
Building AI with pionneers

Stage : Data Scientist (H/F)
Sélection de variables

Qui sommes-nous ?

Nous pourrions commencer en nous présentant comme l'un des leaders sur le marché international du conseil en data, avec un chiffre d'affaires défiant ceux des G.A.F.A. Mais non. Nous préférons mettre en avant notre cadre de travail, nos réussites et surtout nos consultants. Allez venez, c'est par là…

Quantmetry c'est une centaine de personnes qui travaille de concert pour accompagner nos clients dans leurs réflexions, P.O.C., industrialisation autour de la Data, l'Intelligence Artificielle, le Machine Learning et la Big Data. Nous sommes ce qu'on peut appeler communément un cabinet de conseil pure player en Data.
N'étant pas sectorisés, nous intervenons sur différents sujets (voici une petite liste non-exhaustive) :
- La maintenance prédictive,
- La segmentation clients, le scoring,
- La détection de fraude, de corruption, de blanchiment d'argent,
- Le développement des architectures de plateformes Data,
- L'analyse de textes et d'images dans le cadre de la lutte contre le cancer du sein.

Dans le cadre de notre incessante croissance, nous recherchons des personnes motivées pour nous accompagner et ainsi travailler sur des projets de recherche.

Préalable

Quantmetry propose ci-dessous le volet R&D d'une offre de stage. Tout stagiaire entrant chez Quantmetry, outre le travail de R&D qui lui est proposé et pour lequel il sera encadré, aura aussi pour objectif de participer à certaines missions de conseil chez des clients variés, lui permettant d'aborder le monde du consulting.

Quantmetry
Building AI with pionneers

Contexte

La mise en production de modèles de machine learning dans les entreprises se heurte aujourd'hui à plusieurs barrières. Tout d'abord, les modèles sont dits boîtes noires et suscitent la défiance des utilisateurs. Enfin, les modèles sont complexes à mettre à jour car le nombre de variables explicatives est trop élevé et décuple les coûts de maintenance. L'intelligibilité des modèles et leur maintenance en conditions opérationnelles défendent donc la simplicité et le nombre réduit de variables. Les modèles de machine learning sont plus facilement appréhendables par des utilisateurs non-experts, et sont plus faciles à maintenir et mettre à jour, s'ils reposent sur peu de variables. A ce titre, la sélection de variables (feature selection) est un ingrédient primordial de modélisation en machine learning. Il existe beaucoup de méthodes de sélection de variables, mais trop peu sont encore implémentées dans les librairies standard type scikit-learn. D'autres implémentations développées en recherche fondamentale et appliquée existent et sont parfois méconnues de la communauté. Parmi ces implémentations, on trouve en accès libre :

- Stability-selection, développé par QuantumBlack, un cabinet de conseil londonien en Data Science
- MLXtend, développé par l'université du Wisconsin-Madison
- Scikit-feature, développé par l'université d'Arizona

La promesse de ces outils est de fournir une interface type scikit-learn pour diffuser et démocratiser les méthodes de sélection de variables.

Les objectifs de ce stage sont :

- Prendre en main les 3 outils et les appliquer à des jeux de données simples en accès libre
- Comparer leurs fonctionnalités, évaluer leur complétude vis-à-vis de l'état de l'art et lister leurs avantages/inconvénients
- Construire un guide méthodologique et un tutoriel de sélection de variables avec les outils jugés les plus pertinents
- Animer une formation en interne sur la sélection de variables