

Master 2 - ISF

Machine Learning en Python

Consignes du projet

1. Consignes générales

Afin de valider l'UC « Machine Learning en Python », vous devez réaliser un projet (60% de la note finale) par trinômes.

L'objectif de ce projet est de mettre en application la démarche à suivre dans le cadre d'un projet de *Data Science* ainsi que les différents éléments vus en cours (algorithmes, techniques de sampling, cross-validation, one hot encoding, etc.) pour traiter un problème concret (décrit dans §2).

Les livrables attendus de votre part sont :

- **Le code**, dûment commenté, qui vous a permis d'obtenir vos résultats. Ce dernier peut être un Jupyter Notebook ou un script Python réalisé sous un autre logiciel (Spyder, Google Colab, etc.).
- Un **fichier** Jupyter Notebook, HTML, Rmarkdown ou autre **contenant l'exécution de votre code** et affichant vos résultats.
- **Un rapport de maximum 25 pages (au format PDF)** présentant votre démarche, résultats, raisonnement, conclusions et les limites de votre projet. A titre d'exemple, vous pourrez suivre le plan suivant :
 - Présentation/Description des données
 - Retraitements appliqués
 - Modélisation 1
 - Sampling
 - Hyperparamétrage
 - Cross-Validation
 - Résultats
 - Analyse des résultats
 - Modélisation 2
 - ...
 - Conclusion, limites et axes d'améliorations
 - Annexes (non comptabilisés dans les 25 pages).

N'hésitez pas à illustrer vos explications via des captures d'écran et graphiques issus de votre code.

L'envoi du projet doit être fait par un seul membre du groupe (avec les autres membres en CC) et les livrables contenus dans un fichier .zip nommé avec vos noms de famille.

Exemple : le groupe composé de M.X, Mme.Y et M.Z enverra un fichier nommé Projet_X_Y_Z.zip

Les projets doivent être envoyés à l'adresse jorge.ochoa.pro19@gmail.com au plus tard le 17/12/2024 avant minuit.

2. Cas d'usage

Vous êtes un(e) Data Scientist recruté(e) par une banque internationale qui souhaite améliorer son processus de décision pour évaluer le risque associé aux demandes de financement. La banque souhaite améliorer son processus de décision pour évaluer le risque de défaut de paiement avant d'accorder un prêt.

Actuellement, la décision repose sur des règles fixes (par exemple, seuils de revenu ou de ratio dette/revenu), mais ces règles sont trop rigides et ne permettent pas de capturer la complexité des profils clients. Cela entraîne deux problèmes majeurs :

- Acceptation de clients à haut risque, ce qui augmente les pertes liées aux défauts.
- Refus de clients solvables, ce qui réduit les opportunités commerciales.

L'objectif du projet est de développer un modèle de Machine Learning capable de prédire le niveau de risque associé à une demande de prêt à partir des données historiques des clients et de leurs caractéristiques financières. Ainsi, une fois en production votre modèle permettra notamment de :

- Réduire le taux de défaut en identifiant les profils à risque.
- Optimiser le taux d'acceptation en ciblant les clients fiables.
- Automatiser et fiabiliser le processus d'octroi de crédit.

Variables disponibles (cf. le dictionnaire de données pour plus de détails).

Les données fournies proviennent des dossiers passés et incluent :

- **Caractéristiques personnelles** : âge, état civil, nombre de personnes à charge, niveau d'éducation.
- **Situation financière** : revenu, ratio dette/revenu, solde des comptes, actifs et passifs.
- **Historique de crédit** : score de solvabilité, antécédents de faillite, défauts de paiement précédents.
- **Informations sur la demande** : montant, durée, taux d'intérêt.
- **Variable cible** : RiskScore, un indicateur numérique compris entre 0 et 100 reflétant la probabilité de défaut du créancier.

Votre travail consistera à explorer les données, prétraiter les informations, développer et entraîner un ou des modèle(s) de Machine Learning, puis évaluer ses performances. Vous devrez également fournir des recommandations sur la manière dont votre institution peut mettre en œuvre ce modèle de manière efficace pour automatiser le processus d'évaluation du risque.