Livrables et soutenance

Livrables

- L'**API de prédiction** du score, déployée sur le cloud (lien vers l'API).
- Le **notebook ou code** de la modélisation (du prétraitement à la prédiction)
  - Ce notebook intègre la partie MLFlow de génération du tracking d'expérimentations.
  - L'interface web "UI MLFlow" d'affichage des résultats du tracking MLFlow sera présentée en soutenance + copie d'écran dans le support de soutenance.
- Un **dossier,** géré via un outil de versioning de code contenant :
  - Le notebook ou code de la modélisation (du prétraitement à la prédiction), intégrant via MLFlow le tracking d'expérimentations et le stockage centralisé des modèles.
  - Le code permettant de déployer le modèle sous forme d'API.
  - Pour l'API, un fichier introductif permettant de comprendre l'objectif du projet et le découpage des dossiers, et un fichier listant les packages utilisés seront présents dans le dossier.
- Le tableau HTML d'analyse de data drift réalisé à partir d'**evidently.**
- Un **notebook ou une application Streamlit** de test de l'API.
- Un **support de présentation** pour la soutenance, détaillant le travail réalisé (Powerpoint ou équivalent, 30 slides maximum), intégrant des copies écran, preuves qu'un **pipeline de déploiement continu** a permis de déployer l'API :
  - de l'interface web "UI MLFlow" d'affichage des résultats du tracking MLFlow
  - des commits ;
  - du dossier Github (+ lien vers ce dossier) ;
  - de l'exécution des tests unitaires ;
  - de l'exécution du déploiement de l'API avec lien vers l'API sur le Cloud.

Pour faciliter votre passage devant le jury, déposez sur la plateforme, dans un dossier zip nommé ***Titre_du_projet_nom_prénom***, votre livrable nommé comme suit : **Nom_Prénom_n° du livrable_nom du livrable_date de démarrage du projet**. Cela donnera :

- *Nom_Prénom_1_API_mmaaaa*
- *Nom_Prénom_2_notebook_modélisation_mmaaaa*
- *Nom_Prénom_3_dossier_code_mmaaaa*
- *Nom_Prénom_4_Tableau_HTML_data_drift_evidently_mmaaaa*
- *Nom_Prénom_5_notebook_test_API_mmaaaa*
- *Nom_Prénom_6_presentation_mmaaaa*

Par exemple, votre premier livrable peut être nommé comme suit : *Dupont_Jean_1_API_012023.*

Soutenance

Pendant la soutenance, l'évaluateur jouera le rôle de Michaël, à qui vous présentez votre travail.

- **Présentation (20 minutes)**
  - Rappel de la problématique et présentation du jeu de données (2 minutes)
  - Présentation de la modélisation (12 minutes) :
    - Démarche de modélisation, choix des mesures
    - Visualisation du tracking via MLFlow UI
    - Présentation de la synthèse des résultats
    - Analyse de la feature importance globale et locale
  - Présentation du pipeline de déploiement : Git, Github, tests unitaires (2 minutes)
  - Présentation de l'analyse de data drift (2 minutes)
  - Exemple d'un scoring client via appel à l'API sur le Coud (2 minutes)
- **Discussion (5 minutes)**
  - L'évaluateur, jouant le rôle de Michaël, vous challengera sur vos choix.
- **Débriefing (5 minutes)**
  - À la fin de la soutenance, l'évaluateur arrêtera de jouer le rôle de Michaël pour vous permettre de débriefer ensemble.

Votre présentation devrait durer 20 minutes (+/- 5 minutes). Puisque le respect des durées des présentations est important en milieu professionnel, les présentations en dessous de 15 minutes ou au-dessus de 25 minutes peuvent être refusées.

Concernant la mise en production de l'API, plusieurs solutions s'offrent à vous, en particulier Azure webapp, AWS et Heroku. À vous de choisir la solution qui vous convient le mieux.

Dans le cadre de l'utilisation de Heroku, étant devenu payant depuis fin novembre 2022, les coûts liés à votre projet seront **à votre charge**. Vous êtes donc libre de vous investir financièrement si vous le souhaitez, mais vous **n'avez aucune obligation** de le faire pour réaliser ce projet.

Quelque soit la solution Cloud choisie, l'étudiant et l'évaluateur veilleront à enregistrer pendant la soutenance la démo de l'application en production, ce qui permettra au jury de visionner cette démo, sans que l'étudiant n'ait à maintenir son application sur le Cloud. Maintenir l'application dans le Cloud pourrait en effet engendrer des coûts

Référentiel d'évaluation

Définir la stratégie d'élaboration d'un modèle d'apprentissage supervisé et sélectionner et entraîner des modèles adaptés à une problématique métier afin de réaliser une analyse prédictive.

CE1 Les variables catégorielles identifiées ont été transformées en fonction du besoin (par exemple via OneHotEncoder ou TargetEncoder).
CE2 Vous avez a créé de nouvelles variables à partir de variables existantes.
CE3 Vous avez réalisé des transformations mathématiques lorsque c'est requis pour transformer les distributions de variables.
CE4 Vous avez normalisé les variables lorsque c'est requis.
CE5 Vous avez défini sa stratégie d'élaboration d'un modèle pour répondre à un besoin métier. Cela signifie dans ce projet que :

- l'étudiant a présenté son approche méthodologique de modélisation dans son support de présentation pendant la soutenance et est capable de répondre à des questions à ce sujet, si elles lui sont posées.

CE6 Vous avez choisi la ou les variables cibles pertinentes.
CE7 Vous avez vérifié qu'il n'y a pas de problème de data leakage (c'est-à-dire, des variables trop corrélées à la variable cible et inconnues a priori dans les données en entrée du modèle).
CE8 Vous avez testé plusieurs algorithmes de façon cohérente, en partant des plus simples vers les plus complexes (au minimum un linéaire et un non linéaire).

Évaluer les performances des modèles d'apprentissage supervisé selon différents critères (scores, temps d'entraînement, etc.) en adaptant les paramètres afin de choisir le modèle le plus performant pour la problématique métier.

CE1 Vous avez choisi une métrique adaptée pour évaluer la performance d'un algorithme (par exemple : R2 ou RMSE en régression, accuracy ou AUC en classification, etc.). Dans le cadre de ce projet, cela signifie que :

- Vous avez mis en œuvre un score métier pour évaluer les modèles et optimiser les hyperparamètres, qui prend en compte les spécificités du contexte, en particulier le fait que le coût d'un faux négatif et d'un faux positif sont sensiblement différents.

CE2 Vous avez exploré d'autres indicateurs de performance que le score pour comprendre les résultats (coefficients des variables en fonction de la pénalisation, visualisation des erreurs en fonction des variables du modèle, temps de calcul…).
CE3 Vous avez séparé les données en train/test pour les évaluer de façon pertinente et détecter l'overfitting.
CE4 Vous avez mis en place un modèle simple de référence pour évaluer le pouvoir prédictif du modèle choisi (dummyRegressor ou dummyClassifier).
CE5 Vous avez pris en compte dans sa démarche de modélisation l'éventuel déséquilibre des classes (dans le cas d'une classification).
CE6 Vous avez optimisé les hyper-paramètres pertinents dans les différents algorithmes.
CE7 Vous avez mis en place une validation croisée (via GridsearchCV, RandomizedSearchCV ou équivalent) afin d'optimiser les hyperparamètres et comparer les modèles. Dans le cadre de ce projet :

- une cross-validation du dataset train est réalisée ;
- un premier test de différentes valeurs d'hyperparamètres est réalisé sur chaque algorithme testé, et affiné pour l'algorithme final choisi ;
- tout projet présentant un score AUC anormalement élevé, démontrant de l'overfitting dans le GrisSearchCV, sera invalidé (il ne devrait pas être supérieur au meilleur de la compétition Kaggle : 0.82).

CE8 Vous avez présenté l'ensemble des résultats en allant des modèles les plus simples aux plus complexes. Vous avez justifié le choix final de l'algorithme et des hyperparamètres.
CE9 Vous avez réalisé l'analyse de l'importance des variables (feature importance) globale sur l'ensemble du jeu de données et locale sur chaque individu du jeu de données.

Définir et mettre en œuvre un pipeline d'entraînement des modèles, avec centralisation du stockage des modèles et formalisation des résultats et mesures des différentes expérimentations réalisées, afin d'industrialiser le projet de Machine Learning.

CE1 Vous avez mis en oeuvre un pipeline d'entraînement des modèles reproductible.
CE2 Vous avez sérialisé et stocké les modèles créés dans un registre centralisé afin de pouvoir facilement les réutiliser.
CE3 Vous avez formalisé des mesures et résultats de chaque expérimentation, afin de les analyser et de les comparer

Mettre en œuvre un logiciel de version de code afin d'assurer en continu l'intégration et la diffusion du modèle auprès de collaborateurs.

CE1 Vous avez créé un dossier contenant tous les scripts du projet dans un logiciel de version de code avec Git et l'a partagé avec Github.
CE2 Vous avez présenté un historique des modifications du projet qui affiche au moins trois versions distinctes, auxquelles il est possible d'accéder.
CE3 Vous avez tenu à jour et mis à disposition la liste des packages utilisés ainsi que leur numéro de version.
CE4 Vous avez rédigé un fichier introductif permettant de comprendre l'objectif du projet et le découpage des dossiers.
CE5 Vous avez commenté les scripts et les fonctions facilitant une réutilisation du travail par d'autres personnes et la collaboration.

Concevoir et assurer un déploiement continu d'un moteur d'inférence (modèle de prédiction encapsulé dans une API) sur une plateforme Cloud afin de permettre à des applications de réaliser des prédictions via une requête à l'API.

CE1 Vous avez défini et préparé un pipeline de déploiement continu.
CE2 Vous avez déployé le modèle de machine learning sous forme d'API (via Flask par exemple) et cette API renvoie bien une prédiction correspondant à une demande.
CE3 Vous avez mis en œuvre un pipeline de déploiement continu, afin de déployer l'API sur un serveur d'une plateforme Cloud.
CE4 Vous avez mis en œuvre des tests unitaires automatisés (par exemple avec pyTest).
CE5 Vous avez réalisé l'API indépendamment de l'application qui utilise le résultat de la prédiction.

Définir et mettre en œuvre une stratégie de suivi de la performance d'un modèle en production et en assurer la maintenance afin de garantir dans le temps la production de prédictions performantes.

CE1 Vous avez défini une stratégie de suivi de la performance du modèle. Dans le cadre du projet :

- choix de réaliser a priori cette analyse sur le dataset disponible : analyse de data drift entre le dataset train et le dataset test.

CE2 Vous avez réalisé un système de stockage d'événements relatifs aux prédictions réalisées par l'API et une gestion d'alerte en cas de dégradation significative de la performance. Dans le cadre du projet :

- choix de réaliser a priori cette analyse analyse de data drift, via une simulation dans un notebook et création d'un tableau HTML d'analyse avec la librairie **evidently.**

CE3 Vous avez analysé la stabilité du modèle dans le temps et défini des actions d'amélioration de sa performance. Dans le cadre de ce projet :

- analyse du tableau HTML **evidently**, et conclusion sur un éventuel data drift.