Projet N°1 - Advanced Machine Learning

Fall 2020

Date de rendu :
dernier commit le 20 novembre et soumission de la présentation le 23 novembre à 18h

Contact à privilégier :
Sacha Samama (Discord ou sacha@yotta-academy.com)

1. Contexte général

Vous travaillez en tant que consultant Data Scientist (interne ou externe) à l'entreprise Yotta Factory. Vous ainsi que votre équipe, venez d'être affectés au nouveau projet lancé par la Data Factory de votre client. Ce nouveau projet, piloté par le pôle Marketing, est le premier projet que la société souhaite, à terme, industrialiser.

A l'issue du développement, la solution devra être transmise aux équipes métiers (non techniques) qui devront être autonomes pour effectuer leur prédictions.

Le métier attend donc une solution simple à prendre en main, avec des explications et une documentation claires sans avoir à mettre les mains dans le code pour effectuer du debugging.

2. Objectifs de ce projet

Ce projet vise à appréhender "from scratch" une problématique de Data Science en environnement "pré-industrialisé" dans un contexte métier précis et sur des jeux de données que votre équipe vient de recevoir (les cas d'usage et les données sont décrits dans la partie 5.

En plus d'évaluer votre capacité à travailler en équipe, la liste ci-dessous présente l'ensemble des points sur lesquels vous serez évalués pour ce premier projet :

- ❏ Respect des guidelines techniques (décrites dans la partie 3)
- ❏ Savoir mettre en place un pipeline complet de machine learning
- ❏ Savoir restituer et vulgariser l'approche ainsi que la démarche scientifique
- ❏ Fournir un code propre, de qualité et documenté
- ❏ Mobiliser l'ensemble des connaissances et des outils appris jusqu'à présent : Python et son environnement, Gitlab et gestion du code collaboratif, Command Line, Environnements virtuels, Architecture de code et Clean code, EDA, Data Preprocessing, Feature Engineering, Modélisation, Méthode de sélection de modèle,

Optimisation de modèle, Pipeline d'apprentissage et de prédiction, Intelligibilité du modèle, etc.

3. Guidelines techniques à respecter

Ci-dessous, l'ensemble des points à respecter avant de soumettre votre projet :

- Constituer une équipe de 2 à 3 personnes maximum et communiquer votre équipe sous 2 jours.
- Créer un repository Gitlab sous ce groupe gitlab.
- Respecter l'arborescence du Workflow Gitlab ou *Gitlab flow*
- Chaque membre de l'équipe devra avoir réalisé un minimum de 3 commits et 1 merge sur une des branches principales du projet.
- Le code devra être modulaire, documenté et bien architecturé (template DDD recommandé).
- Le traitement des données devra être développé façon *Pipeline() scikit-learn.*
- Le code devra être fonctionnel et facilement rejouable sur un environnement quelconque (celui du métier) → packagé
- Un sous ensemble des jeux de données ont été conservés pour évaluer la facilité de prise en main de la solution (run des prédictions et restitution des résultats).
- Faire de belles data visualisation claires
- Le code devra contenir un pipeline d'apprentissage ET un pipeline de prédiction.
- Le projet devra être documenté (à minima un README) : cette documentation facilitera la prise en main de votre solution par l'équipe métier.
- Support de présentation clair et concis

4. Format de restitution

3 supports de restitution sont attendus :

1. Le projet packagé poussé dans un repository Gitlab et dans lequel la branche principale à jour sera taguée "*v1*".

2. Le projet devra contenir 3 notebooks nettoyés illustrant : la *phase exploratoire*, recherche d'un modèle optimal (via *Grid Search)* et la phase d'intelligibilité. Ces notebooks devront utiliser le package (fonctions, classes, méthodes, etc.) que vous avez développé.

3. Un support de présentation détaillant :
    a. L'ensemble des phases de la problématique : résultat de l'analyse exploratoire, compréhension du problème, démarche scientifique complète, transformation

des données, choix et critères de sélection du modèle retenu, réflexion sur l'intelligibilité du modèle développé

b. Ainsi que les points suivants : fonctionnement de l'équipe, répartition des tâches et outils et framework utilisés (IDE, environnement virtuel, librairies, organisation du code et du git, etc.)

Ces supports seront évalués lors d'une soutenance de présentation de 20 minutes qui aura lieu le 20 novembre 2020 sur un créneau qui sera précisé. Une séance de questions-réponses de 15 minutes suivra la présentation.

Note : positionnez-vous en tant qu'expert s'adressant à un public de non initié à la data science.

5. Cas d'usage et description des données

Trois problématiques sont à votre disposition ; votre groupe devra en sélectionner une au choix et se référer au dictionnaire de données associés.

Sujet 1 : Churn Modelling

→ Prédiction de l'attrition client au sein d'une banque.

Sujet 2 : Lead Scoring

→ Prédiction de prospects à fort potentiel pour un organisme de formation.

Sujet 3 : Product Subscription

→ Prédiction de l'appétence des clients à un produit bancaire.