Projet N°1 - Advanced Machine Learning

Avril 2020

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

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

1. Contexte général

Vous travaillez en tant que consultant Data Scientist indépendant ou bien êtes une ressource interne au sein d'une équipe data indépendante dans l'entreprise X. Vous et votre équipe venez d'être affectés au nouveau projet lancé par la X Data Factory. Ce nouveau projet, initié par le pôle Marketing, constitue le premier projet que l'entité souhaite industrialiser rapidement.

A l'issue du développement, la solution devra être transmise aux équipes métiers marketing (non techniques) qui devront être autonomes pour effectuer leur prédictions. Ils ne devront pas vous solliciter pour comprendre pourquoi la ligne 100 de votre code "a crashé".

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

2. Objectifs de ce projet

Ce projet vise à appréhender "from scratch" une problématique de Data Science dans un contexte métier précis et sur un jeu de données que votre équipe vient de recevoir (les données sont décrites en 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 (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 et de qualité
- ❏ Mobiliser l'ensemble des connaissances et des outils appris jusqu'à présent : Python et son environnement, Gitlab et gestion du code collaboratif, Command Line, Architecture de code, Clean code, Analyse exploratoire, Data Cleaning &

Preprocessing, Feature Engineering, Modélisation, Méthode de sélection de modèle, Optimisation de modèle, Pipeline d'apprentissage et de prédiction, Intelligibilité du modèle retenu, etc.

3. Guidelines 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 code devra être fonctionnel et facilement rejouable sur un environnement quelconque (celui du métier).
- 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).
- Les notebooks ne sont permis que pour les phases d'exploration et d'intelligibilité.
- 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
- **BONUS : implémenter vos propres Pipeline() scikit-learn.**

4. Format de restitution

2 supports de restitution sont attendus :

1. Le code packagé poussé dans un repository Gitlab et dans lequel la branche principale à jour sera taguée "*v1*" (ie. la branche "v1" devra apparaître sur Gitlab).

2. 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 24 avril 2020 sur un créneau à préciser. 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. Description des données

Les données sont tirées de campagnes marketing direct d'une institution bancaire. Ces campagnes de marketing direct sont basées sur des appels téléphoniques. Souvent, plusieurs contacts avec un même client sont nécessaires pour savoir si le produit (dépôt à terme) sera ou non souscrit.

Note : un **dépôt à terme** est un dépôt qu'une banque ou une institution financière offre à un taux fixe (souvent mieux que l'ouverture d'un compte de dépôt) et dans lequel votre argent vous sera rendu à une échéance précise.

Le dossier ***data/*** contient deux jeux de données complémentaires mis à disposition : *data.csv* et *socio_demo.csv.* Ci-dessous la description des 2 jeux de données :

Note : ces deux jeux de données sont stockés à l'identique au sein du SI du Datalab.

socio_eco.csv

Jeu de données contenant des indicateurs socio-économique, contexte de marché.

Champs	Description
DATE	Date de fin de mois
EMPLOYMENT_VARIATION_RATE	Taux de variation de l'emploi. Indicateur trimestriel
IDX_PRICE_CONSUMER	Indice des prix à la consommation. Indicateur mensuel
IDX_CONSUMER_CONFIDENCE	Indice de confiance des consommateurs. Indicateur mensuel
NB_EMPLOYE	Nombre d'employés. Indicateur trimestriel

data.csv

Jeu de données principal

Champs	Description
DATE	Date du dernier contact
AGE	Age en années
JOB_TYPE	Type de métier
STATUS	Statut marital
EDUCATION	Niveau de diplôme
HAS_DEFAULT	Le client a déjà fait défaut ou non
BALANCE	Balance du client
HAS_HOUSING_LOAN	Le client détient un prêt immobilier
HAS_PERSO_LOAN	Le client détient un prêt conso
CONTACT	Type de communication du contact
DURATION_CONTACT	Durée du dernier contact, en secondes. Note : cet attribut affecte la cible de sortie (par exemple, si = 0 alors SUBSCRIPTION = “No”). Cependant, la durée n'est pas connue avant l'exécution d'un appel. De même, après la fin de l'appel, SUBSCRIPTION est évidemment connu.
NB_CONTACT	Nombre de contacts effectués pendant cette campagne et pour ce client (inclut le dernier contact)
NB_DAY_LAST_CONTACT	Nombre de jours écoulés après que le client a été contacté pour la dernière fois lors d'une campagne précédente (999 signifie que le client n'a pas été contacté auparavant)
NB_CONTACT_LAST_CAMPAIGN	Nombre de contacts effectués avant cette campagne et pour ce client
RESULT_LAST_CAMPAIGN	Résultat de la précédente campagne marketing
SUBSCRIPTION	Le client a-t-il souscrit ou non