Projet #3

Prédiction du Churn Client avec des Modèles de Machine Learning

Introduction au Machine Learning

Architecte en Intelligence Artificielle - Année 1 - **L'École Multimédia**

Sommaire

Contexte

Présentation

Vous êtes embauché en tant data scientiste au sein d'une entreprise de télécommunications.

Votre mission consiste à développer un modèle de Machine Learning capable de prédire le churn des clients, c'est-à-dire le risque qu'ils résilient leur abonnement.

Vous devrez explorer les données des clients, sélectionner et entraîner différents modèles de régression et de classification, et optimiser ces modèles pour obtenir les meilleures prédictions possibles.

Vous présenterez également vos résultats sous forme de visualisations claires et d'une documentation détaillée.

Contraintes

- **Travail en autonomie** : Vous travaillerez seul sur ce projet, ce qui implique une gestion efficace du temps et des ressources.
- **Documentation complète** : Chaque étape du projet doit être clairement documentée
- **Utilisation de Git et GitHub** : Tout le projet doit être versionné sur GitHub, avec des commits réguliers et bien nommés (tel que https://www.conventionalcommits.org), un README détaillé, et une bonne structuration du code.
- **Respect des délais** : Le projet doit être livré sous la forme d'une **archive ZIP** et dans les délais imposés sous peine de pénalisation.

Libertés

- **Bibliothèques** : Dans la mesure où la base de votre code est écrit en Python, vous pouvez utiliser les bibliothèques (apprentissage, calcul numérique, visualisation) qui vous sembleront les plus appropriées ;
- **Algorithmes** : Vous pouvez utiliser les algorithmes que vous souhaitez (à condition de le justifier) ;
- **Déploiement** : Bien que cela ne soit pas le cœur du projet, vous pouvez proposer des pistes pour le déploiement en ligne de votre modèle, via un tableau de bord de consultation des données, soit via une API (avec Flask ou Fast API, par exemple) ;
- **Améliorations** : En fonction de votre familiarité avec le sujet, vous pouvez proposer des options qui permettraient d'améliorer le pipeline de base de l'apprentissage.

Objectifs

Vous devrez avoir réalisé les éléments suivant :

1. Collecter et préparer les données nécessaires pour prédire le churn des clients.
2. Développer et comparer différents modèles de Machine Learning (régression logistique, arbres de décision, Random Forest) pour identifier les clients à risque.
3. Évaluer et optimiser les performances des modèles en utilisant des techniques de validation et d'optimisation des hyperparamètres.
4. Visualiser les résultats pour aider à interpréter les prédictions et faciliter la prise de décision.
5. Documenter le processus complet et présenter les conclusions à l'équipe dirigeante.

Étapes du projet

- Collecte et Préparation des Données
  - Collecte des données : Utilisez un dataset public (par exemple, le dataset "Telco Customer Churn" de Kaggle) ou des données internes fournies par l'entreprise.
  - Préparation des données : Nettoyez les données, traitez les valeurs manquantes, encodez les variables catégorielles, et normalisez les variables numériques pour améliorer les performances des modèles.
- Exploration des Données
  - Analyse exploratoire des données (EDA) : Analysez les caractéristiques des clients qui sont restés et ceux qui ont *churné*, en utilisant des visualisations avec Seaborn et Matplotlib.
  - Sélection des caractéristiques : Identifiez les variables les plus pertinentes pour prédire le churn, telles que l'âge, le type de contrat, le nombre de services souscrits, etc.
- Développement des Modèles de Régression
  - Régression Logistique :
    - Implémentez un modèle de régression logistique pour prédire le churn.
    - Évaluez les performances du modèle à l'aide de matrices de confusion, courbes ROC, et scores AUC.
- Comparaison avec d'autres modèles de régression :
  - Testez d'autres modèles de régression pour déterminer si la régression logistique est la plus adaptée.
- Développement des Modèles de Classification Avancés
  - Arbres de Décision :
    - Construisez un arbre de décision pour modéliser la probabilité de churn en fonction des caractéristiques des clients.
    - Interprétez les règles de décision et analysez la complexité du modèle.

        - Random Forest :
            - Implémentez un modèle de Random Forest pour améliorer la précision des prédictions.
            - Analysez l'importance des caractéristiques et optimisez les hyperparamètres du modèle.
- Évaluation et Optimisation des Modèles
    - Comparaison des modèles : Comparez les performances des différents modèles en utilisant des métriques telles que la précision, le rappel, le score F1, et le score AUC.
    - Optimisation des hyperparamètres : utilisez des techniques comme la validation croisée et la recherche en grille pour ajuster les hyperparamètres et maximiser les performances.
- Visualisation et Présentation des Résultats
    - Visualisation des performances : Créez des graphiques avec Matplotlib et Seaborn pour illustrer les performances des modèles (courbes ROC, importance des caractéristiques, etc.).
    - Création de dashboards : utilisez Plotly ou Bokeh pour créer des visualisations interactives permettant d'explorer les prédictions de churn.
    - Préparation d'une présentation : Rédigez un rapport et préparez une présentation pour expliquer les résultats et les recommandations basées sur ces résultats.
- Documentation et Gestion du Projet
    - Documentation : Documenter chaque étape du processus, incluant la préparation des données, l'implémentation des modèles, et l'analyse des résultats.
    - Utilisation de GitHub : Gérez votre projet avec Git et GitHub, en versionnant le code source, en suivant les modifications, et en collaborant avec vos coéquipiers.

Données

Pour réaliser les objectifs du projet, vous utiliserez le jeu de données que vouspourrez télécharger via la ligne de commande suivante :

```
#!/bin/bash

curl -L -o ~/Downloads/telco-customer-churn.zip
https://www.kaggle.com/api/v1/datasets/download/blastchar/telco-customer-churn
```

Contenu

Chaque ligne représente un client, chaque colonne contient les attributs du client décrits dans la colonne Métadonnées.

L'ensemble de données comprend des informations sur :

- Les clients qui ont quitté dans le dernier mois – la colonne est appelée Churn

- Les services auxquels chaque client s'est abonné – téléphone, Internet, sécurité en ligne, sauvegarde en ligne, protection de l'appareil, support technique, et streaming TV et films
- Informations sur le compte client – combien de temps ils ont été client, contrat, mode de paiement, facturation sans papier, frais mensuels et frais totaux
- Informations démographiques sur les clients – sexe, tranche d'âge, et si elles ont des partenaires et des personnes à charge

Rendu final

Votre rendu final prendra la forme d'une archive Zip et devra comporter les éléments suivants :

1. Un document (comme un carnet Jupyter, par exemple) dans lequel vous analyserez le jeu de données fourni et où vous préparerez le ce jeu de données de manière à ce qu'il soit acceptable comme entrée de votre processus d'apprentissage. Dont :
    a. Une analyse de la qualité de données
    b. Une analyse statistique descriptive des données (avec visualisation)
    c. Une procédure pour nettoyer le jeu de données
2. Un document développant la mise en œuvre du processus d'apprentissage, en détaillant et justifiant :
    a. Le choix d'un algorithme particulier
    b. Les hyperparamètres du modèle
    c. Les indicateurs de performance de votre modèle
3. une procédure de validation du modèle, sur un jeu de données de test, avec une analyse de la capacité de votre modèle à généraliser son apprentissage et les conclusions que vous en tirez.

Cette archive aura comme titre votre nom et prénom avec votre classe.

Exemple: `arthur_mensch_projet3_AIA01.zip`

Attention : Un rendu non livré ou en retard vous pénalise pour la certification

Evaluations

- Qualité de la préparation des données : Pertinence du nettoyage, de la normalisation, et de la sélection des caractéristiques.
- Performance des modèles développés : Précision des prédictions et efficacité des techniques de modélisation utilisées.
- Capacité à évaluer et optimiser les modèles : Utilisation appropriée des métriques et des techniques d'optimisation pour améliorer les performances des modèles.
- Clarté et interactivité des visualisations : Efficacité des visualisations pour transmettre les résultats et soutenir la prise de décision.
- Documentation et gestion du projet : Exhaustivité et clarté de la documentation, ainsi que l'utilisation appropriée de GitHub pour la gestion du projet.
- Présentation professionnelle des résultats : Qualité de la présentation finale, incluant l'explication des choix méthodologiques, l'interprétation des résultats, et les recommandations.

Compétences à valider

D-02	Créer un algorithme d'Intelligence Artificielle adapté aux données d'entraînement et conforme aux spécifications du cahier des charges, en veillant à répondre aux besoins spécifiques, notamment en termes d'accessibilité.
D-04	Concevoir des pipelines d'intégration et déploiement continu pour automatiser le processus de déploiement d'une solution d'IA.
D-06	Piloter la performance de la solution d'IA dans l'infrastructure à travers la mise en place d'outils de monitoring (comme Aporia ou Evidently) pour s'assurer qu'elle respecte les spécifications du cahier des charges dans un environnement de production.

Conseils

- Bien prendre le temps d’analyser le brief et comprendre le client
- Organisez-vous et planifiez votre travail : donnez vous des objectifs intermédiaires
- Planifiez des sessions de travail régulière
- Utilisez Git pour versionner votre code dès le départ
- Ne jamais être trop ambitieux
- Faites directement les documents du livrable
- Mettez en oeuvre les bonnes pratiques vues en cours
- Refactoriser pour éviter le code redondant
- Soignez la qualité de votre code (commentaires, indentation)
- Pensez à la qualité du résultat !