DataSpace

Plan prévisionnel

Dataset retenu

Le dataset (Lending Club Loan Data) retenu contient des informations sur des prêts accordés sur une période s'étendant de 2007 à 2016. Il comprend environ 42 538 observations et 52 variables. Les données fournissent des détails sur l'état actuel du prêt (tel que Current, Late, Fully Paid, etc.), les informations de paiement les plus récentes, ainsi que des caractéristiques telles que les cotes de crédit, le nombre de demandes de financement, l'adresse (y compris les codes postaux et les États), et d'autres informations pertinentes pour l'évaluation des risques de crédit et la gestion des prêts. Un dictionnaire de données est disponible pour fournir une description détaillée de chaque variable.

Modèle envisagé

L'algorithme NGBoost est choisi pour son efficacité et sa pertinence dans le cadre du projet, qui spécifie explicitement de choisir un modèle datant de moins de cinq ans. NGBoost répond à cette exigence, ayant été introduit pour la première fois en 2019. Cette fraîcheur assure que l'algorithme intègre les avancées technologiques les plus récentes en matière d'apprentissage automatique, le plaçant au premier plan en termes de performance et de pertinence pour les problèmes contemporains.
De plus, des études récentes ont démontré que NGBoost surpasse de nombreux autres algorithmes d'apprentissage automatique en termes de performance, en particulier dans des tâches de classification et de régression. Son adaptabilité, sa flexibilité et sa capacité à maintenir une certaine interprétabilité en font un choix judicieux pour diverses applications, de la classification de crédit à la détection de fraude.
Ainsi, en choisissant NGBoost, nous nous assurons de répondre aux exigences du projet tout en exploitant un modèle de pointe capable de fournir des résultats précis et interprétables pour résoudre les défis de modélisation actuels.

Références bibliographiques

1. Chen, T., Guestrin, C. (2019). NGBoost: Natural Gradient Boosting for Probabilistic Prediction. arXiv preprint arXiv:1910.03225. [Article de recherche]
    - Cet article présente NGBoost, un nouvel algorithme de boosting de gradient qui améliore les performances de prédiction probabiliste. Il fournit une explication détaillée de l'algorithme et démontre son efficacité à travers des expériences sur différents ensembles de données.
2. Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media. [Livre]

- Ce livre est une ressource précieuse pour comprendre les concepts de base de l'apprentissage automatique ainsi que pour apprendre à utiliser les bibliothèques Python populaires comme Scikit-Learn, Keras et TensorFlow. Il propose des exemples pratiques et des cas d'utilisation qui peuvent aider à approfondir la compréhension du sujet.

3. Brownlee, J. (2021). "NGBoost for Probabilistic Classification and Regression in Python." Machine Learning Mastery. [Blog post]
   - Cet article de blog fournit un guide pratique sur l'utilisation de NGBoost pour la classification et la régression probabilistes en Python. Il explique étape par étape comment mettre en œuvre l'algorithme avec des exemples de code et offre des conseils pour optimiser ses performances.

.

Explication de votre démarche de test du nouvel algorithme (votre preuve de concept)

Pour tester le nouvel algorithme NGBoost, nous allons suivre une approche en deux étapes :

1. **Baseline avec des modèles traditionnels** :
   - Tout d'abord, nous allons établir une baseline en utilisant des modèles traditionnels tels que la régression logistique, les arbres de décision et les forêts aléatoires. Ces modèles serviront de référence pour évaluer les performances de NGBoost. Nous utiliserons des métriques standard telles que l'accuracy, le score Jaccard et la perte de Hamming pour comparer les résultats.

2. **Implémentation de NGBoost** :
   - Ensuite, nous mettrons en œuvre l'algorithme NGBoost sur les mêmes données et évaluerons ses performances par rapport à la baseline établie précédemment. Nous ajusterons les hyperparamètres du modèle pour optimiser ses performances et utiliserons les mêmes métriques d'évaluation pour la comparaison.

Dans le cadre de la data science et du machine learning, cette approche de preuve de concept nous permettra de déterminer si NGBoost est viable pour notre problème spécifique et d'évaluer sa performance par rapport aux méthodes traditionnelles. En utilisant un ensemble de données limité, nous pourrons obtenir des résultats rapidement et décider si une exploration plus approfondie de NGBoost est justifiée pour résoudre notre problème.