Note méthodologique : preuve de concept

Dataset retenu

Dans le cadre de ce projet, nous avons utilisé le dataset (Lending Club Loan Data) qui porte sur des prêts accordés entre 2007 et 2016. Il comprend des informations détaillées sur l'état actuel de chaque prêt ainsi que les dernières données de paiement. Ce fichier contient environ 42 538 observations et 52 variables. Les données incluent une variété de caractéristiques démographiques et financières des emprunteurs, offrant ainsi une vue complète sur les différents aspects des prêts et des emprunteurs.

Parmi les informations cruciales, on trouve le statut du prêt, qui indique l'état actuel du prêt, Cela peut être soit "Fully Paid" lorsque le prêt a été entièrement remboursé, soit "Charged Off" lorsqu'il est considéré comme une perte car l'emprunteur ne remboursera probablement pas, ou bien d'autres états tels que "Current" pour les prêts en cours ou "Late" pour les prêts en retard de paiement. Cette variable est essentielle pour comprendre la situation des prêts au moment de l'analyse et pour évaluer le risque associé à chaque prêt.

Le dataset comprend également des informations de paiement, notamment les dernières données sur les paiements effectués par les emprunteurs. Ces données permettent d'analyser les tendances de remboursement et d'identifier les facteurs qui influencent la régularité des paiements. En outre, les cotes de crédit des emprunteurs sont incluses, fournissant une mesure clé pour évaluer la solvabilité des individus.

Les caractéristiques des emprunteurs incluent le nombre de demandes de financement soumises par les emprunteurs, ce qui peut être un indicateur de la fréquence des besoins de financement et de la dépendance à l'égard du crédit. Les informations géographiques, telles que les adresses, les codes postaux et les états de résidence des emprunteurs, permettent d'effectuer des analyses régionales et de comprendre les variations géographiques dans les comportements de prêt.

En plus des informations démographiques et géographiques, le dataset contient des données sur les collections et l'historique de paiement, offrant une vue d'ensemble sur les antécédents financiers des emprunteurs. Ces informations sont essentielles pour une évaluation complète du risque de crédit.

Le dictionnaire de données fourni dans un fichier distinct décrit chaque variable en détail, aidant ainsi les analystes à comprendre le rôle et l'importance de chaque élément du dataset.

Les concepts de l’algorithme récent

L'algorithme Gradient Boosting (GBoost) repose sur le principe de combiner de manière séquentielle de nombreux modèles simples, appelés "weak learners", pour former un modèle plus puissant. Cette approche itérative permet de corriger les erreurs des modèles précédents en se concentrant sur les observations les plus difficiles à prédire.

Le processus débute par l'initialisation du modèle avec une prédiction naïve, souvent la moyenne de la variable cible pour les problèmes de régression, ou la classe majoritaire pour les problèmes de classification.

À chaque étape itérative, un nouveau modèle est ajouté pour améliorer les prédictions du modèle existant. Ce nouveau modèle est formé sur les résidus des prédictions précédentes. Les modèles utilisés sont généralement des arbres de décision peu profonds, également appelés "trees". Le taux d'apprentissage, qui contrôle la contribution de chaque modèle, est ajusté pour favoriser la convergence du processus d'apprentissage.

Les prédictions de tous les modèles sont ensuite agrégées en sommant les prédictions pondérées de chaque modèle. La pondération de chaque modèle dépend de son taux d'apprentissage et de sa performance sur les résidus.

Le processus de formation se poursuit jusqu'à ce qu'un critère d'arrêt soit atteint, tel qu'un nombre maximal d'itérations ou une convergence des résidus.

Les avantages de l'algorithme GBoost sont nombreux. Il offre une grande flexibilité, une précision élevée et une interprétabilité. Il est capable de s'adapter à différents types de données et de problèmes, produisant des prédictions précises même dans des ensembles de données complexes. Bien que le modèle global soit complexe, les modèles individuels (arbres de décision) sont souvent faciles à interpréter, ce qui permet de comprendre les relations entre les variables et de tirer des insights pertinents.

La modélisation

La méthodologie de modélisation adoptée repose sur plusieurs étapes clés pour développer un modèle prédictif robuste et précis.

1. **Exploration des Données**: Cette étape consiste à explorer et à comprendre les données disponibles. Cela inclut l'analyse des distributions des variables, la détection des valeurs aberrantes et la corrélation entre les différentes caractéristiques.
2. **Prétraitement des Données**: Avant de construire le modèle, il est nécessaire de nettoyer et de prétraiter les données. Cela peut impliquer le traitement des valeurs manquantes, la normalisation des variables et la conversion des variables catégorielles en format adapté au modèle.
3. **Choix du Modèle**: Selon la nature du problème (régression, classification, etc.) et les caractéristiques des données, plusieurs types de modèles peuvent être envisagés. Pour ce projet, l'algorithme Gradient Boosting (GBoost) peut être une option pertinente

compte tenu de sa capacité à gérer des ensembles de données complexes et à produire des prédictions précises.

4. **Division des Données**: Les données sont divisées en ensembles d'entraînement et de test. L'ensemble d'entraînement est utilisé pour ajuster le modèle, tandis que l'ensemble de test est utilisé pour évaluer sa performance.
5. **Construction du Modèle**: Le modèle GBoost est construit en utilisant l'ensemble d'entraînement. Différents hyperparamètres peuvent être ajustés pour optimiser les performances du modèle, tels que le taux d'apprentissage, la profondeur des arbres, et le nombre d'itérations.
6. **Évaluation du Modèle**: Une fois le modèle construit, il est évalué sur l'ensemble de test à l'aide de métriques d'évaluation appropriées pour le type de problème (par exemple, RMSE pour les problèmes de régression, et l'accuracy, precision, recall pour les problèmes de classification).
7. **Optimisation du Modèle**: Si nécessaire, des techniques d'optimisation peuvent être utilisées pour améliorer les performances du modèle. Cela peut inclure l'ajustement des hyperparamètres, la sélection de variables, ou même l'exploration de différentes architectures de modèle.
8. **Validation Croisée**: Pour garantir la robustesse du modèle, une validation croisée peut être effectuée. Cela implique la division des données en plusieurs sous-ensembles et l'évaluation du modèle sur chaque sous-ensemble de manière itérative.

Métrique d'Évaluation et Démarche d'Optimisation

Dans le cas de ce projet, nous avons utilisé pour l'évaluation de nos modèles des métriques telles que l'accuracy, precision, recall et F1-score peuvent être utilisées pour évaluer la performance du modèle.

La démarche d'optimisation implique l'ajustement des hyperparamètres du modèle pour maximiser sa performance. Cela peut être réalisé à l'aide de techniques d'optimisation telles que la recherche par grille, la recherche aléatoire, ou même des méthodes plus avancées telles que l'optimisation bayésienne.

Synthèse des résultats

La synthèse des résultats comparés entre les techniques utilisées précédemment et la technique récente montre des performances remarquables de la technique récente, notamment le modèle NGBoost.

Résultats des Techniques Précédentes

Les techniques précédentes comprenaient l'utilisation de modèles tels que la Régression Logistique, l'Arbre de Décision et la Forêt Aléatoire. Les résultats obtenus avec ces modèles ont été les suivants :

- **Régression Logistique** :
    - Précision : 97.4%
    - Jaccard Score : 97.1%
    - Hamming Loss : 2.6%
- **Arbre de Décision** :
    - Précision : 99.8%
    - Jaccard Score : 99.8%
    - Hamming Loss : 0.2%
- **Forêt Aléatoire** :
    - Précision : 99.9%
    - Jaccard Score : 99.9%
    - Hamming Loss : 0.1%

Ces résultats sont excellents, mais ils ont été obtenus à partir de techniques traditionnelles qui peuvent être limitées par leur capacité à capturer des modèles complexes dans les données.
Résultats de la Technique Récente (NGBoost)
La technique récente utilisée est le modèle NGBoost. Les résultats obtenus avec ce modèle sont les suivants :

- **NGBoost** :
    - Précision : 99.9%
    - Jaccard Score : 99.8%
    - Hamming Loss : 0.1%

Le modèle NGBoost affiche des performances comparables voire supérieures à celles des techniques précédentes. Il offre une précision et un score Jaccard élevés tout en maintenant un faible taux de perte. De plus, l'utilisation de NGBoost permet une meilleure interprétabilité grâce à la possibilité d'obtenir les importances des caractéristiques.

L’analyse de la feature importance globale et locale du nouveau modèle

L'analyse de l'importance des caractéristiques (feature importance) globale et locale du nouveau modèle NGBoost permet de comprendre quelles caractéristiques ont le plus d'impact sur les prédictions globales du modèle et comment ces caractéristiques influent sur les prédictions individuelles.

Importance des Caractéristiques Globale

L'importance globale des caractéristiques nous indique l'impact relatif de chaque variable sur les prédictions du modèle à l'échelle de l'ensemble des données. Cette analyse permet d'identifier les caractéristiques les plus importantes pour le modèle dans son ensemble.
D'après l'analyse de l'importance globale des caractéristiques avec NGBoost, nous pouvons observer les caractéristiques les plus influentes sur les prédictions globales du modèle. Par exemple, les caractéristiques telles que le montant du prêt (loan_amnt), le taux d'intérêt (int_rate), le montant de l'acompte (installment), et le montant total payé (total_pymnt) peuvent avoir une influence significative sur les décisions de prêt.

Importance des Caractéristiques Locale
L'importance locale des caractéristiques, parfois appelée feature contribution ou SHAP values, nous fournit des informations sur l'impact spécifique de chaque caractéristique pour chaque prédiction individuelle. Cette analyse nous permet de comprendre comment chaque caractéristique contribue à la prédiction pour des cas spécifiques.
En examinant l'importance locale des caractéristiques avec NGBoost, nous pouvons voir comment chaque caractéristique influence les prédictions pour des prêts individuels. Par exemple, nous pouvons constater que pour certains prêts, le montant du prêt ou le taux d'intérêt peuvent être des facteurs déterminants dans la décision d'octroi du prêt, tandis que pour d'autres prêts, d'autres caractéristiques telles que le revenu annuel ou le nombre d'années d'expérience peuvent jouer un rôle plus important.

Les limites et les améliorations possibles

Limites de l'Approche Actuelle

1. **Dépendance aux Données** :

L'approche actuelle de modélisation dépend fortement des données disponibles. Si les données sont biaisées ou incomplètes, cela peut affecter la performance du modèle et introduire des biais dans les prédictions.

2. **Interprétabilité Limitée** :

Bien que le modèle NGBoost offre de bonnes performances prédictives, son interprétabilité peut être limitée. Comprendre comment les caractéristiques influent sur les prédictions reste complexe, ce qui peut rendre difficile l'explication des décisions du modèle.

3. **Complexité du Modèle** :

Le modèle NGBoost est complexe et peut nécessiter des ressources de calcul importantes pour l'entraînement et l'inférence, surtout avec de grandes quantités de données. Cela peut limiter sa scalabilité et sa capacité à être déployé dans des environnements avec des contraintes de ressources.

Améliorations Envisageables

1. **Gestion des Données Déséquilibrées** :

Une amélioration possible consiste à développer des techniques de gestion des données déséquilibrées plus sophistiquées, en particulier pour les problèmes de classification où les classes sont déséquilibrées. L'utilisation de méthodes d'échantillonnage plus avancées ou de techniques de pondération peut permettre d'améliorer la performance du modèle.

2. **Interprétabilité Renforcée** :

Pour améliorer l'interprétabilité du modèle, des techniques telles que l'utilisation de SHAP (SHapley Additive exPlanations) values ou d'autres méthodes d'exploration de l'importance des caractéristiques peuvent être employées. Cela permettrait de mieux comprendre les décisions du modèle et d'expliquer ses prédictions de manière plus intuitive.

3. **Optimisation des Hyperparamètres** :

Une autre amélioration possible consiste à effectuer une optimisation plus approfondie des hyperparamètres du modèle NGBoost. Cela peut inclure l'utilisation de techniques d'optimisation bayésienne ou de recherche en grille pour trouver les meilleurs paramètres du modèle et améliorer sa performance.

4. **Enrichissement des Données** :

En enrichissant les données avec des informations supplémentaires pertinentes, telles que des variables dérivées ou des données externes, on peut potentiellement améliorer la capacité du modèle à capturer des modèles plus complexes et à fournir des prédictions plus précises.

5. **Déploiement sur des Plateformes Légères** :

Pour rendre le modèle plus facilement déployable dans des environnements avec des contraintes de ressources, il est possible d'explorer des techniques de réduction de la taille du modèle ou de déploiement sur des plateformes légères telles que des dispositifs mobiles ou des services cloud à faible latence.