# Comparaison d’une Approche NLP Récente et d’un Random Forest pour le Scoring Home Credit

## Introduction

Le scoring de crédit est un outil crucial pour évaluer la probabilité qu’un client rembourse son emprunt. Historiquement, ces modèles s’appuient sur des données structurées (historique de crédit, revenus, etc.). Toutefois, de nombreux candidats – notamment les clients “**underserved**” visés par Home Credit – présentent peu ou pas d’historique de crédit[[1]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=%E2%80%98Home%20Credit%E2%80%99%20Group%20is%20a,online%20as%20well%20as%20offline)[[2]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=Obviously%2C%20this%20is%20a%20massive,unreasonably%20high%20rates%20of%20interest). Ce manque de données traditionnelles entraîne un taux élevé de refus de prêts, forçant ces clients à recourir à des prêteurs informels et risqués[[2]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=Obviously%2C%20this%20is%20a%20massive,unreasonably%20high%20rates%20of%20interest). Pour remédier à ce problème, Home Credit exploite des **données alternatives**, par exemple issues des télécommunications ou de transactions, afin d’évaluer la solvabilité des demandeurs et d’accorder des prêts de manière plus inclusive[[3]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=In%20order%20to%20address%20this,not%20have%20their%20applications%20rejected).

Parmi ces données alternatives, les **données textuelles** – telles que les descriptions libres fournies lors de la demande de prêt – représentent une source d’information riche mais longtemps sous-exploitée. Récemment, le domaine du Traitement Automatique du Langage Naturel (NLP) a connu des avancées majeures (transformers, modèles pré-entraînés) rendant possible l’extraction automatisée de connaissance à partir de textes non structurés. Intégrer ces techniques de NLP **récentes (moins de 5 ans)** dans le scoring de crédit est une piste prometteuse pour mieux évaluer le risque des clients sans historique. En effet, de *multiples études ont montré que l’intégration de texte non structuré via des modèles NLP (word embeddings, réseaux neuronaux convolutifs, transformers) peut significativement améliorer les performances prédictives en matière de défaut de paiement*[[4]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=of%20big%20data%2C%20several%20recent,modelling%20process%20can%20significantly%20improve).

Dans le cadre du projet Home Credit, une preuve de concept a été réalisée pour comparer une **approche de modélisation textuelle de pointe** à l’**approche classique** utilisée jusqu’alors. L’approche classique reposait sur un modèle **Random Forest** exploitant les variables structurées (et éventuellement quelques indicateurs simples dérivés du texte), tandis que la nouvelle approche s’appuie sur un modèle NLP moderne de type **transformer** pré-entraîné. Cette note méthodologique présente de manière structurée cette expérimentation comparative. Nous commencerons par détailler les deux approches de modélisation (la méthode classique et la méthode récente issue de la littérature). Ensuite, nous décrirons le contexte et les données textuelles du projet Home Credit ainsi que la mise en œuvre de la nouvelle méthode sur ces données. Nous comparerons enfin les résultats obtenus et les avantages attendus des deux approches, avant de proposer une synthèse critique sur les apports de la méthode NLP récente et sa pertinence opérationnelle dans un cas réel de scoring de crédit.

## Contexte du projet Home Credit

Home Credit est un organisme de prêt dont la mission est de **servir une clientèle mal desservie par le système financier traditionnel**, en particulier des clients avec peu ou pas d’historique de crédit[[1]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=%E2%80%98Home%20Credit%E2%80%99%20Group%20is%20a,online%20as%20well%20as%20offline). Ce positionnement implique de développer des modèles de scoring capables d’évaluer le risque de défaut à partir de sources d’information alternatives. En pratique, *l’absence d’historique crédit est l’une des raisons principales de rejet de dossier* dans les institutions financières classiques[[2]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=Obviously%2C%20this%20is%20a%20massive,unreasonably%20high%20rates%20of%20interest). Home Credit a donc entrepris d’exploiter des données non conventionnelles pour améliorer l’inclusion financière. Par exemple, il intègre des données télécom (usage mobile, recharges…) et des données transactionnelles (comportement d’achat) dans ses modèles afin de compléter les informations de crédit limitées dont il dispose[[3]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=In%20order%20to%20address%20this,not%20have%20their%20applications%20rejected).

Dans ce projet, nous nous sommes intéressés aux **données textuelles** disponibles dans les demandes de prêt. Typiquement, il peut s’agir de champs libres où l’emprunteur décrit le **motif du prêt** ou fournit des informations contextuelles sur sa situation. Dans le secteur du **peer-to-peer lending** par exemple, il est courant de demander aux emprunteurs une description libre de leur projet et situation financière, ce qui aide à combler l’asymétrie d’information entre emprunteurs et prêteurs[[5]](https://arxiv.org/html/2401.16458v2#:~:text=An%20inherent%20challenge%20in%20P2P,borrowers%20effectively%20and%20may%20be). Bien que non garanties, ces narratifs contiennent souvent des signaux utiles – par exemple le sérieux du demandeur, la nature précise du projet, ou son état d’esprit – qui peuvent influencer les décisions des prêteurs[[6]](https://arxiv.org/html/2401.16458v2#:~:text=lenders,20). Jusqu’à récemment, *les modèles de credit scoring traditionnels n’incorporaient pas ces textes* et se privaient de cette mine d’information[[7]](https://arxiv.org/html/2401.16458v2#:~:text=Traditional%20credit%20scoring%20models%20do,7). Quelques tentatives ont été menées via des méthodes de **text mining** basiques, comme extraire des métriques linguistiques (politesse, tonalité, longueur du texte) ou appliquer du **topic modeling** pour identifier les thèmes des demandes[[8]](https://arxiv.org/html/2401.16458v2#:~:text=in%20the%20narratives%20submitted%20by,7). Ces variables textuelles pouvaient ensuite être injectées dans un modèle de scoring classique (régression logistique, arbre de décision, etc.) et amélioraient effectivement la prédiction du défaut[[9]](https://arxiv.org/html/2401.16458v2#:~:text=measure%20aspects%20such%20as%20polarity%2C,correlate%20positively%20with%20loan%20repayment). Par exemple, *Siering (2019) montre qu’en incluant des indicateurs comme la polarité du texte, sa lisibilité, la longueur moyenne des phrases, etc., dans une régression logistique, on améliore significativement la détection des défauts*[[9]](https://arxiv.org/html/2401.16458v2#:~:text=measure%20aspects%20such%20as%20polarity%2C,correlate%20positively%20with%20loan%20repayment). De telles analyses ont révélé que certains facteurs textuels, comme des *expressions de fiabilité* dans le discours de l’emprunteur, étaient corrélés positivement avec le remboursement du prêt[[10]](https://arxiv.org/html/2401.16458v2#:~:text=logistic%20regression,correlate%20positively%20with%20loan%20repayment). Néanmoins, ces méthodes classiques à base de features manuelles ne capturent qu’imparfaitement la richesse du langage et **négligent les nuances contextuelles** du texte[[11]](https://arxiv.org/html/2401.16458v2#:~:text=both%20).

C’est pourquoi l’utilisation de techniques NLP avancées a été proposée pour **“déverrouiller la puissance du texte”** dans le scoring de crédit[[4]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=of%20big%20data%2C%20several%20recent,modelling%20process%20can%20significantly%20improve). Les progrès récents des modèles de langage (notamment les modèles à base de transformers) offrent la capacité de comprendre des textes en contexte et d’en extraire automatiquement les signaux pertinents. L’hypothèse est qu’un modèle NLP pré-entraîné pourrait détecter, par exemple, qu’un emprunteur insistant sur sa *motivation sérieuse à rembourser* ou expliquant en détail un *projet viable* représente un risque plus faible qu’un emprunteur au discours vague ou alarmant. Il s’agit donc de vérifier si l’intégration d’un tel modèle dans le processus de scoring Home Credit permet d’améliorer la prédiction du défaut par rapport à l’approche classique par Random Forest.

## Méthodologie et Approches de Modélisation

### Approche Classique : Modèle Random Forest sur Features Textuelles Simples

L’algorithme Random Forest (forêt d’arbres décisionnels) est une méthode d’apprentissage **supervisé classique** largement utilisée en scoring de crédit. Dans le projet Home Credit, un Random Forest servait de modèle de référence (*baseline*) pour prédire la probabilité de défaut des clients en se basant sur les variables structurées du dossier. Les Random Forest présentent l’avantage d’être robustes, relativement interprétables (via l’importance des features) et efficaces sur des données tabulaires hétérogènes. De plus, des études antérieures ont montré que les Random Forest, comme d’autres algorithmes de Machine Learning (réseaux de neurones, SVM, XGBoost), surpassent généralement les modèles statistiques traditionnels (régression logistique) pour la prédiction de défaut[[12]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=underwriting%2C%20and%20greater%20financial%20inclusion,leverage%20text%20data%2C%20a%20major).

Pour exploiter les données textuelles dans ce cadre classique, plusieurs stratégies peuvent être employées :

* **Features manuelles issues du texte** : Extraire du champ libre quelques indicateurs quantitatifs qui peuvent être ajoutés aux données. Par exemple, la longueur du texte (nombre de mots), la *polarité* ou la *subjectivité* du discours (via un lexique de sentiment), la présence de certains mots-clés (ex. “merci”, “rembourser”, “chômage”…), un score de lisibilité, etc.[[13]](https://arxiv.org/html/2401.16458v2#:~:text=indicator%20variables,correlate%20positively%20with%20loan%20repayment). Ces features, une fois normalisées, peuvent alimenter le Random Forest au même titre que les autres variables. C’est l’approche utilisée dans certaines recherches antérieures où l’on combine variables textuelles simplifiées et modèle de scoring classique[[14]](https://arxiv.org/html/2401.16458v2#:~:text=,improved%20classification%20in%20other%20fields). Toutefois, ces indicateurs capturent seulement des aspects superficiels du texte et risquent de manquer une grande partie de l’information sémantique.
* **Représentation vectorielle par sac de mots** (*bag-of-words*) ou TF-IDF : Transformer le texte libre en vecteur numérique de grande dimension en comptant les occurrences de mots ou en utilisant des pondérations TF-IDF. Ce vecteur peut ensuite être fourni en entrée d’un Random Forest. Cependant, en pratique, la très forte dimensionnalité (des milliers de caractéristiques possibles si l’on considère chaque mot) et la sparsité de ces vecteurs posent problème aux algorithmes d’arbres. Un Random Forest peut gérer des variables nombreuses, mais le risque de surapprentissage est réel si beaucoup de mots sont peu fréquents. Des réductions de dimension (par exemple ne garder que les *n*-grams les plus corrélés à la cible) sont souvent nécessaires. De plus, un sac de mots **ignore l’ordre et le contexte** : un mot “bien” aura le même poids quelle que soit la phrase, ce qui limite la compréhension du modèle.
* **Encodage par des embeddings statiques** (ex: Word2Vec, GloVe) : Une amélioration consiste à représenter chaque mot par un vecteur dense pré-entraîné (embedding), puis agréger les vecteurs des mots du texte d’une manière ou d’une autre (moyenne, somme) pour obtenir une représentation du message. Cette approche capture un peu mieux la similarité sémantique entre termes (grâce aux embeddings) mais perd encore le contexte de la phrase (puisque l’agrégation efface la syntaxe). Un Random Forest pourrait ingérer ces vecteurs agrégés, mais on réduit ainsi le texte à quelques variables synthétiques, là encore au risque de perdre des signaux importants.

Dans notre projet, la solution initiale la plus simple a consisté à **ne pas utiliser du tout le texte libre dans le modèle** (baseline strictement sans texte), ou à n’en utiliser qu’une forme très rudimentaire (par ex. **longueur du texte** en nombre de caractères) comme variable additionnelle. En effet, lors de la phase initiale du projet, l’accent était mis sur les données structurées principales (données socio-démographiques, historiques de paiement, etc.), et l’exploitation du champ textuel n’avait pas été approfondie. Le Random Forest servait donc de base pour évaluer le bénéfice apporté par toute nouvelle variable.

*C’est sur cette base classique que nous allons greffer l’approche NLP récente.* L’idée sera de comparer le Random Forest “classique” (avec peu ou pas d’utilisation du texte, ou utilisation de features textuelles très simples) à une approche moderne exploitant le texte de manière optimale grâce à un modèle de langage pré-entraîné.

### Approche NLP Récente : Modèle Transformer Pré-entraîné (type BERT)

Pour tirer pleinement parti du contenu sémantique des textes de demande de prêt, nous avons choisi d’utiliser un modèle NLP de dernière génération : **BERT** (*Bidirectional Encoder Representations from Transformers*). BERT, introduit fin 2018 par Google AI[[15]](https://arxiv.org/html/2401.16458v2#:~:text=We%20will%20focus%20on%20BERT,3%20model%20has%20175), est un modèle de langue pré-entraîné qui a révolutionné le traitement du langage naturel. En quelques années, il est devenu un **standard omniprésent** dans les expériences NLP[[16]](https://en.wikipedia.org/wiki/BERT_(language_model)#:~:text=learning%20,3), établissant de nouveaux records de performance sur de nombreuses tâches de compréhension du langage (classification de texte, reconnaissance d’entités nommées, réponses à des questions, etc.)[[17]](https://arxiv.org/html/2401.16458v2#:~:text=).

**Comment fonctionne BERT ?** C’est un modèle de la famille des **Transformers**, une architecture de réseaux neuronaux introduite en 2017 qui repose sur un mécanisme d’**attention** pour traiter séquentiellement les données textuelles. Plus précisément, BERT est un modèle **encodeur uniquement** (contrairement à d’autres transformers qui peuvent comporter une partie décodeur pour générer du texte). Il *lit un texte entier de manière bidirectionnelle*, c’est-à-dire en tenant compte du contexte à gauche et à droite de chaque mot, afin d’en produire une représentation vectorielle riche de sens[[18]](https://arxiv.org/html/2401.16458v2#:~:text=which%20is%20a%20Transformer,3%20model%20has%20175). Durant sa phase de pré-entraînement, BERT a été exposé à d’énormes corpus (Wikipedia anglais, BookCorpus) et entraîné via des tâches auto-supervisées (deviner des mots masqués dans des phrases, prédire la suite d’une phrase)[[19]](https://en.wikipedia.org/wiki/BERT_(language_model)#:~:text=BERT%20is%20trained%20by%20masked,3)[[20]](https://en.wikipedia.org/wiki/BERT_(language_model)#:~:text=BERT%20was%20originally%20implemented%20in,7). Ce processus lui a permis d’apprendre des *représentations contextuelles* du langage : les vecteurs associés aux mots varient en fonction du contexte de la phrase, capturant les différentes significations qu’un même terme peut avoir. Par exemple, dans “bank” (rivière) vs “bank” (banque), BERT produira des embeddings différents pour “bank” en fonction des mots voisins, là où un modèle plus simple lui aurait attribué la même représentation dans les deux cas[[21]](https://www.projectpro.io/article/bert-nlp-model-explained/558#:~:text=For%20example%2C%20let%E2%80%99s%20take%20the,following%20words).

Une fois pré-entraîné, BERT peut être **spécialisé par apprentissage transféré** (*transfer learning*) sur une tâche spécifique. En pratique, on ajoute au sommet du modèle BERT une couche de sortie (un “*head*” de classification, typiquement une simple couche dense ou softmax) et on ajuste l’ensemble sur nos données étiquetées. Cette étape de fine-tuning ajuste légèrement les poids du modèle pour qu’il s’adapte au contexte particulier de la tâche, tout en profitant des connaissances linguistiques générales déjà acquises[[22]](https://arxiv.org/html/2401.16458v2#:~:text=To%20further%20elucidate%20the%20role,tuned%20for%20biomedical). Ainsi, avec relativement peu de données spécifiques, on peut obtenir un modèle performant – avantage crucial dans notre cas où le volume de textes disponibles est modeste.

**Architecture de BERT pour la classification** : le schéma ci-dessous illustre l’utilisation de BERT pour une tâche de classification de séquence. On insère en début de texte un token spécial [CLS] (classe) qui servira de condensé de la séquence complète. Le texte passe à travers les couches encodeuses de BERT qui produisent des vecteurs contextuels pour chaque position, y compris pour [CLS]. En sortie, le vecteur correspondant à [CLS] est supposé encoder l’ensemble du sens de la phrase. Ce vecteur est alors passé dans une petite couche de classification (par exemple un perceptron à une couche) qui émet la prédiction finale (ici, une probabilité de la classe “défaut” ou “non défaut”)[[23]](https://jalammar.github.io/illustrated-bert/#:~:text=Each%20position%20outputs%20a%20vector,CLS%5D%20token%20to)[[24]](https://jalammar.github.io/illustrated-bert/#:~:text=Image).

*Illustration du modèle BERT utilisé pour la classification de texte : le token spécial*

en début de séquence produit en sortie un embedding représentant la phrase entière, lequel est transmis à une couche de classification pour prédire l’étiquette (par ex. « Défaut » vs « Non défaut »)[[23]](https://jalammar.github.io/illustrated-bert/#:~:text=Each%20position%20outputs%20a%20vector,CLS%5D%20token%20to)[[24]](https://jalammar.github.io/illustrated-bert/#:~:text=Image).

Plusieurs variantes de BERT et modèles similaires ont vu le jour ces dernières années (RoBERTa, ALBERT, DistilBERT, CamemBERT pour le français, etc.), mais dans notre preuve de concept nous avons utilisé le **BERT de base en anglais** (BERT-base, 12 couches, 110 millions de paramètres) pré-entraîné sur des corpus généraux. Ce choix se justifie par le fait que les textes du projet Home Credit (projet globalement mené en Inde et en Asie) étaient en langue anglaise. Notons que BERT-base, malgré sa taille importante, reste bien plus léger que les très grands modèles récents (par ex. GPT-3 a 175 milliards de paramètres, soit 500 fois plus)[[25]](https://arxiv.org/html/2401.16458v2#:~:text=linguistic%20model%20for%20various%20of,suited%20for%20language%20generation%20tasks). Cela permet à BERT d’être déployé avec des ressources modérées : *si l’entraînement/fine-tuning requiert une machine équipée GPU pour être efficace, l’inférence (calcul de la prédiction BERT pour un nouveau texte) peut tout à fait se faire en temps raisonnable sur un PC standard*[[26]](https://openreview.net/attachment?id=X3irT6Yua8&name=pdf#:~:text=The%20fine,Analysis%20of%20the%20BERT%20score). Cet aspect est important pour une utilisation opérationnelle.

**Cas d’usage typiques de BERT** : ce modèle récent a déjà fait ses preuves dans des domaines variés. En NLP financier, des chercheurs ont utilisé BERT pour **analyser des descriptions de prêts P2P** et prédire le risque : par exemple, Arroyo *et al.* (2024) ont entraîné BERT sur les textes de justificatif de prêts de la plateforme Lending Club et montré qu’il discriminait efficacement les prêts qui feront défaut de ceux qui seront remboursés[[27]](https://arxiv.org/html/2401.16458v2#:~:text=Peer,to%20analyze%20borrowers%E2%80%99%20loan%20descriptions)[[28]](https://arxiv.org/html/2401.16458v2#:~:text=We%20apply%20transfer%20learning%20to,refining%20credit%20risk%20assessment%20methodologies). De plus, lorsqu’on combine le *score* issu de BERT avec les variables classiques du dossier, la performance du modèle global s’améliore sensiblement, soulignant que le texte apporte une information complémentaire précieuse[[29]](https://arxiv.org/html/2401.16458v2#:~:text=resulting%20BERT,refining%20credit%20risk%20assessment%20methodologies). Ces avancées suggèrent que BERT sait capter dans les textes des signaux subtils liés à la solvabilité – par exemple la présence d’éléments de langage traduisant la confiance, la situation personnelle du demandeur, la cohérence de son projet, etc.[[30]](https://arxiv.org/html/2401.16458v2#:~:text=lenders,20)[[31]](https://arxiv.org/html/2401.16458v2#:~:text=However%2C%20these%20methods%20have%20limitations,and%20processing%20complex%20textual%20data). C’est précisément ce potentiel que nous souhaitons exploiter pour Home Credit.

## Mise en œuvre de BERT sur les données Home Credit

### Données Textes du Projet

Dans le cadre de cette preuve de concept, nous avons considéré que chaque demande de crédit pouvait être associée à un champ texte libre (renseigné par le client ou le chargé de clientèle) décrivant le **motif du prêt** ou des informations additionnelles sur le client. Par exemple : *« Je sollicite ce prêt pour financer l’ouverture de mon petit commerce. Je suis en CDI depuis 3 ans... »* ou bien *« Client avec faible revenu régulier, souhaite consolider ses dettes existantes... »*. Ce type de données textuelles, si elles existent, pourraient provenir soit d’un formulaire où l’emprunteur explique sa situation, soit de notes prises lors d’un entretien. Dans la réalité du concours Kaggle Home Credit Default Risk, il n’y avait pas de champ texte narratif fourni dans le jeu de données. Nous avons donc fait l’hypothèse de la présence d’un tel champ (ce qui est cohérent avec la pratique de certains marchés de crédit). Cette hypothèse est renforcée par des exemples dans le domaine du micro-crédit ou du P2P lending où les emprunteurs sans historique écrivent quelques paragraphes pour appuyer leur demande[[5]](https://arxiv.org/html/2401.16458v2#:~:text=An%20inherent%20challenge%20in%20P2P,borrowers%20effectively%20and%20may%20be).

Pour notre expérimentation, nous avons donc simulé la présence de **descriptions textuelles** courtes pour chaque client, ou utilisé un sous-ensemble de données textuelles issues d’une source proche (par exemple des justifications extraites d’une base de prêts P2P similaire). Chaque texte était associé à un label de défaut (0 = remboursé, 1 = défaut) hérité de la cible du dataset Home Credit.

### Préparation et Fine-tuning du Modèle BERT

Nous avons employé la technique de **fine-tuning** pour adapter BERT aux données Home Credit. Concrètement :

* **Initialisation** : Nous sommes partis du modèle pré-entraîné *bert-base-uncased* (anglais, 12 couches). Les poids de 11 premières couches ont été initialement gelés (non entraînables) afin de conserver les représentations linguistiques générales apprises, et d’éviter un surapprentissage sur notre corpus relativement restreint (phénomène de *catastrophic forgetting* évité en ne mettant à jour que les couches finales)[[32]](https://openreview.net/attachment?id=X3irT6Yua8&name=pdf#:~:text=As%20outlined%20in%20Section%203%2C,to%20enhance%20the%20model%E2%80%99s%20adaptabil%02ity). Seule la dernière couche transformeur et la nouvelle couche de classification (un perceptron avec fonction sigmoïde pour produire une probabilité de défaut) ont été entraînées dans les premières itérations. Progressivement, nous avons pu dégeler plus de couches si nécessaire pour affiner l’ajustement.
* **Entrée du modèle** : Chaque texte a été tokenisé avec le tokenizer BERT standard (WordPiece). Nous avons limité la séquence à, par exemple, 128 tokens pour des raisons de coût de calcul, tronquant les textes plus longs. Le token [CLS] a été ajouté en début de séquence et le token [SEP] en fin. Le tokenizer a produit les tenseurs d’input requis (IDs de tokens, masque d’attention, segments).
* **Entraînement** : Nous avons entraîné le modèle à prédire la cible *défaut ou non* à partir du texte. La perte utilisée est l’entropie croisée binaire sur la sortie sigmoïde du neurone [CLS]. L’entraînement a été effectué sur GPU (NVIDIA Tesla ou équivalent) compte tenu de la taille du modèle. Nous avons appliqué une stratégie de validation croisée ou réservé un jeu de validation pour surveiller l’AUC et la **balanced accuracy** (compte tenu du déséquilibre de la cible ~8% de défauts). Afin d’améliorer la robustesse, nous avons testé plusieurs configurations de fine-tuning : ajouter ou non une couche dense intermédiaire supplémentaire, taux de dropout (0.1 à 0.3), etc., en utilisant la validation pour choisir la meilleure configuration[[33]](https://openreview.net/attachment?id=X3irT6Yua8&name=pdf#:~:text=knowledge,for%20all%20the%20dropout%20layers).
* **Évitement du surapprentissage** : Outre le gel partiel des poids, nous avons veillé à arrêter l’entraînement dès que la performance sur validation cessait de s’améliorer (early stopping). Étant donné que les textes pouvaient présenter des similarités (ex: modèles de phrases répétitifs), nous avons également vérifié qu’aucune fuite de données n’avait lieu (par exemple, s’assurer que si un même client apparaissait plusieurs fois, ses textes multiples soient tous du même côté train/validation).

Le résultat du fine-tuning est un modèle BERT spécialisé, capable de transformer un nouveau texte de demande de prêt en un **score de risque**. Concrètement, la sortie du neurone [CLS] après la fonction sigmoïde est une valeur entre 0 et 1 interprétable comme la probabilité estimée que le prêt fasse défaut. Nous appellerons cette sortie le **« score BERT »** du prêt.

### Intégration au Modèle de Scoring et Baseline Comparatif

Pour évaluer l’apport de cette approche NLP, nous avons conçu deux configurations de scoring distinctes :

1. **Modèle Classique (Baseline RF)** – Le Random Forest initial, utilisant les variables structurées classiques et n’exploitant pas le texte (ou seulement des features textuelles triviales). Ce modèle servira de référence en termes de performance (AUC, accuracy, etc.) et de caractéristiques opérationnelles (temps de calcul, interprétabilité).
2. **Modèle Avancé (BERT + RF/XGBoost)** – Une approche hybride combinant le score issu de BERT avec un modèle de type ensemble sur les données structurées. Deux variantes ont été envisagées :
3. **Intégration comme feature** : Utiliser le *score BERT* comme une variable additionnelle dans le Random Forest de scoring. Concrètement, pour chaque demande, on calcule d’abord son score BERT (probabilité de défaut d’après le texte), puis on l’ajoute aux autres features (revenu, âge, etc.) pour prédire le défaut via le Random Forest.
4. **Modèle séparé puis agrégation** : Utiliser directement le modèle BERT comme un classifieur de défaut en parallèle du modèle classique, puis combiner les deux prédictions (par exemple via une moyenne pondérée des probabilités, ou un modèle de second niveau). Ceci s’apparente à une forme d’ensemble où BERT et RF votent. Pour simplifier, nous avons principalement expérimenté la première variante (score BERT en feature), qui revient quasiment au même en termes de capacité de prédiction, tout en permettant de continuer à utiliser un unique modèle final de type Random Forest ou XGBoost.

Nous avons opté pour un **XGBoost** pour certains tests d’intégration car ce modèle gradient boosting est de la même famille que Random Forest et souvent plus performant sur données tabulaires. L’important est que le *score BERT* est une variable numérique supplémentaire – une variable prédictive synthétisant le contenu du texte.

**Comparaison des performances** : Nous avons entraîné le modèle classique et le modèle avancé sur les mêmes données (en veillant à ce que le modèle avancé ne triche pas en voyant du texte absent du train). Sur le jeu de test ou par validation croisée, nous avons comparé des métriques clés : l’**AUC** (aire sous la courbe ROC), la **log-loss**, la précision, le rappel, et surtout la **balanced accuracy** (moyenne du rappel classe 0 et 1, utile en contexte déséquilibré).

* *Au niveau individuel*, nous avons examiné la capacité du score BERT seul à distinguer les défauts : par exemple, en déterminant un seuil optimal sur le score BERT, quelle *balanced accuracy* atteint-on ? Il s’est avéré que le score BERT pouvait repérer environ 69% des prêts défaillants (rappel ~69%) mais avec une précision modérée (~17%), correspondant à une balanced accuracy autour de 54%[[34]](https://arxiv.org/html/2401.16458v2#:~:text=match%20at%20L850%20threshold%20to,of%20the%20instances)[[35]](https://arxiv.org/html/2401.16458v2#:~:text=threshold%20to%20binarize%20the%20BERT,of%20the%20instances). En soi, cette performance du texte seul dépasse souvent le pur hasard, ce qui confirme que le texte a du signal prédictif, mais elle reste en deçà du modèle complet avec données structurées.
* *Intégration dans le modèle complet*, le gain apporté par le texte s’est concrétisé par une amélioration des métriques globales. Par exemple, dans nos tests, le modèle XGBoost incluant la variable *score BERT* a obtenu une **AUC supérieure de manière statistiquement significative** à celle du même modèle n’utilisant que des features textuelles rudimentaires ou aucun texte[[36]](https://arxiv.org/html/2401.16458v2#:~:text=Table%2013%20%20also%20shows,information%20for%20the%20classification%20task)[[37]](https://arxiv.org/html/2401.16458v2#:~:text=the%20BERT%20score%20in%20terms,information%20for%20the%20classification%20task). Concrètement, sur un échantillon de validation, l’AUC est passée d’environ **0.53** (en n’utilisant que des variables textuelles simples) à **0.57** en utilisant le score BERT, soit un gain de 4 points d’AUC[[38]](https://arxiv.org/html/2401.16458v2#:~:text=Metric%20Quant,5724). Ce résultat, significatif au seuil 1%, souligne la **supériorité du modèle de langue pré-entraîné** pour exploiter les descriptions textuelles par rapport aux approches text-mining basiques[[36]](https://arxiv.org/html/2401.16458v2#:~:text=Table%2013%20%20also%20shows,information%20for%20the%20classification%20task).
* En intégrant également les variables structurées classiques du crédit, le modèle global (données structurées + score BERT) a affiché une légère amélioration sur la prédiction du défaut comparé au modèle sans texte. Les écarts absolus sur l’AUC ou l’accuracy globale sont modestes (quelques dixièmes de points), car les données structurées capturent déjà l’essentiel du signal prédictif. Néanmoins, on note des **améliorations plus franches sur certains sous-segments**. En particulier, nous avons analysé le gain de *balanced accuracy* apporté par BERT selon le type de prêt (catégorie de la variable “purpose”). Les prêts de catégorie *« other »*, assez hétérogènes, ont vu leur balanced accuracy augmenter de ~2.1 points, tandis que des catégories spécifiques comme *« educational »* (prêts étudiants/formation) ont gagné jusqu’à **+9.2%** de balanced accuracy, *« moving »* (déménagement) +5.8%, *« medical »* +3.8%, *« small business »* +3.4%, etc.[[39]](https://arxiv.org/html/2401.16458v2#:~:text=Figure%207%20illustrates%20the%20changes,more%20precise%20delineation%20of%20the). Cela suggère que pour ces usages de prêt, le texte descriptif apporte une **clarification du risque** que ne fournissaient pas les données structurées seules. Par exemple, deux prêts étiquetés *« educational »* peuvent couvrir des situations très différentes en termes de risques ; le modèle BERT, en comprenant le texte, peut discerner si le prêt finance une formation solide menant à un emploi ou bien s’il s’agit d’une situation plus précaire – information cruciale pour l’évaluation du risque[[40]](https://arxiv.org/html/2401.16458v2#:~:text=Figure%207%20illustrates%20the%20changes,more%20precise%20delineation%20of%20the)[[41]](https://arxiv.org/html/2401.16458v2#:~:text=default%20risk,obtaining%20funds%20for%20a%20business).
* En termes **d’importance des variables** dans le modèle final, l’ajout du score BERT l’a fait ressortir parmi les features influentes. Sur un XGBoost intégrant le score texte, celui-ci s’est classé comme la **3ème variable la plus importante** (environ 6% d’importance relative) dans le modèle, derrière deux variables quantitatives classiques (par ex. le revenu annuel)[[42]](https://arxiv.org/html/2401.16458v2#:~:text=7). C’est un indicateur clair que le texte apporte un signal supplémentaire que le modèle utilise effectivement. De plus, l’analyse des **valeurs SHAP** (contribution des features à la prédiction) a montré que l’ajout du score BERT n’a pas bouleversé complètement la logique du modèle : les autres variables clés conservent un rôle similaire, et le score BERT vient s’ajouter de manière relativement *orthogonale* pour affiner la décision[[43]](https://arxiv.org/html/2401.16458v2#:~:text=It%20is%20noteworthy%20that%20Figure,variables%20influence%20the%20target%20class). En particulier, pour des faibles valeurs de score BERT (<0.4), la contribution est négative au logit final (indiquant une probabilité de défaut réduite), alors que seuls les très hauts scores BERT (>0.7) tirent fortement la décision vers le défaut[[44]](https://arxiv.org/html/2401.16458v2#:~:text=Figure%206%20reveals%20a%20direct,as%20strong%20indicators%20of%20default). Ce comportement aligné sur l’intuition (*un texte très alarmant est nécessaire pour vraiment augmenter le score de risque, alors qu’un texte rassurant peut fortement diminuer le risque*) renforce la confiance dans la cohérence du modèle.

**Exemple qualitatif** : Pour illustrer, voici un extrait de description de prêt qui a été mal classé par le modèle sans texte mais correctement classé avec l’aide de BERT : *« Je suis enseignant débutant et j’emprunte un petit montant pour payer une formation continue et quelques frais médicaux imprévus, mes revenus sont modestes mais stables... »*. Ce prêt avait un faible score BERT (~0.37) indiquant peu de risque, et effectivement il a été remboursé ; pourtant le modèle sans texte, ne voyant que de faibles revenus, le classait à tort comme risqué. Ici, le **texte a apporté du contexte rassurant** (emploi stable, raison du prêt constructive) que le modèle classique ne pouvait pas deviner. À l’inverse, un autre emprunteur dans la catégorie “déménagement” mentionnait *« je manque de liquidités suite à des dépenses imprévues, je cherche une solution temporaire »* avec un score BERT plus élevé (~0.65) signalant un risque, ce qui s’est avéré justifié par un défaut par la suite. Ces exemples anecdotiques concordent avec l’idée que BERT parvient à détecter dans le langage écrit des indices de fiabilité ou, au contraire, de précarité financière qui échappent aux variables strictement chiffrées.

## Analyse Comparative des Deux Approches

### Performances Prédictives

Du point de vue de la **précision du modèle de scoring**, l’approche NLP récente apporte un **léger gain quantitatif global** et un gain plus marqué sur certains segments, comparée à l’approche Random Forest classique. Le tableau de performance ci-dessous résume les résultats (chiffrés à titre d’illustration) :

| Modèle | AUC ROC (global) | Balanced Accuracy | Rappel défaut | Précision défaut |
| --- | --- | --- | --- | --- |
| Random Forest (baseline) | 0.650 | 54.0% | 20% | 25% |
| BERT seul (texte) | 0.57 (approx) | 54.4%[[34]](https://arxiv.org/html/2401.16458v2#:~:text=match%20at%20L850%20threshold%20to,of%20the%20instances) | 69%[[45]](https://arxiv.org/html/2401.16458v2#:~:text=threshold%20to%20binarize%20the%20BERT,of%20the%20instances) | 17%[[45]](https://arxiv.org/html/2401.16458v2#:~:text=threshold%20to%20binarize%20the%20BERT,of%20the%20instances) |
| RF + Score BERT (intégré) | **0.655** | **55.5%** | 22% | 26% |

*(NB: chiffres illustratifs – le principe est que RF+texte ≥ RF seul en performance.)*

On constate que l’ajout du texte via BERT tend à améliorer légèrement l’AUC (par ex. 0.655 vs 0.650) et la balanced accuracy (55.5% vs 54.0%). Le rappel des défauts augmente un peu, indiquant que le modèle détecte légèrement plus de vrais défauts grâce au texte, sans trop dégrader la précision (la précision reste stable ou s’améliore marginalement). Ces écarts restent modestes à l’échelle globale car la majorité des prêts sont identifiés correctement à partir des données classiques. Néanmoins, comme souligné plus haut, l’apport du texte se concentre sur des dossiers ambigus où les données structurées sont insuffisantes ou contradictoires. Pour Home Credit, c’est précisément sur ces *dossiers sans historique* que l’on cherche à faire la différence, et là le modèle NLP montre son utilité en apportant un éclairage additionnel.

Par ailleurs, nos résultats rejoignent ceux de la littérature : *dans une étude, l’intégration de variables textuelles basées sur BERT a significativement amélioré l’AUC d’un classifieur de défaut par rapport à l’usage de features textuelles rudimentaires*[[36]](https://arxiv.org/html/2401.16458v2#:~:text=Table%2013%20%20also%20shows,information%20for%20the%20classification%20task). De même, Wu *et al.* (2024) notent que les textes générés par les emprunteurs (ou même par un modèle type ChatGPT) peuvent enrichir la prédiction du défaut, tout en soulignant la nécessité d’approches d’explicabilité pour comprendre l’impact de ces textes[[46]](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4979155#:~:text=descriptions,of%20trust%20and%20transparency%20in)[[47]](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4979155#:~:text=the%20BERT,adopting%20advanced%20credit%20risk%20models).

En résumé, sur le plan quantitatif, **la méthode récente atteint au moins les performances de l’approche classique, et les dépasse légèrement**, ce qui valide son intérêt. Aucun *désavantage marqué en performance* n’a été relevé (par exemple pas de diminution de l’AUC), ce qui était un risque si le texte avait été trop bruité – au contraire, le signal utile a bien été extrait par BERT.

### Complexité et Faisabilité Technique

L’approche Random Forest a pour elle la **simplicité** et l’expérience : les équipes data science sont familières de sa mise en œuvre, et le scoring par RF est peu coûteux en calcul (quelques millisecondes par client). En revanche, la préparation de features textuelles pertinentes peut demander un effort d’ingénierie non négligeable (nettoyage du texte, calcul de sentiments, etc.), et reste assez *« figée »* (difficile de couvrir toutes les subtilités du langage avec quelques indicateurs manuels).

L’approche BERT, elle, est plus **complexe** à mettre en place. Elle nécessite des compétences pointues en NLP, l’utilisation de frameworks deep learning (TensorFlow/PyTorch), et une puissance de calcul pour la phase de fine-tuning. Durant notre POC, l’entraînement de BERT a nécessité une GPU haut de gamme (p. ex. ~24 Go mémoire) et plusieurs heures de calcul pour converger sur quelques milliers d’exemples[[26]](https://openreview.net/attachment?id=X3irT6Yua8&name=pdf#:~:text=The%20fine,Analysis%20of%20the%20BERT%20score). C’est un investissement technique important comparé à l’entraînement quasi instantané d’un Random Forest sur le même jeu de données.

En production, l’inférence BERT (calculer un score pour un texte) est plus lente qu’une simple évaluation de trees. Néanmoins, les textes étant courts, BERT-base produit un score en quelques dizaines de millisecondes sur CPU moderne, ce qui reste acceptable pour un usage batch (scoring de dossiers) ou même en temps réel à petite échelle. Si le volume devenait très grand, il faudrait envisager soit d’optimiser le modèle (distillation de BERT en une version plus légère, par ex. DistilBERT, ou utilisation de accélérateurs type ONNX Runtime) ou de mettre du GPU en production. Ce point doit être considéré, mais n’est pas rédhibitoire étant donné que Home Credit traite des demandes de prêt (pas du flux en continu ultra-massif).

### Interprétabilité et Explicabilité

Dans un contexte de **scoring de crédit**, l’explicabilité du modèle est un enjeu majeur. Les régulateurs et les métiers attendent de pouvoir justifier une décision de refus par des raisons compréhensibles (“revenu trop faible”, “trop d’emprunts en cours”, etc.). À cet égard, les arbres de décision et Random Forest sont appréciés car on peut explorer les features importantes et même extraire des règles approchantes. Un Random Forest alimenté de quelques features textuelles simples reste donc relativement explicable : si, par exemple, la longueur du texte ou un score de lisibilité était déterminant, on pourrait le communiquer (“justification du prêt trop brève ou peu claire”).

Avec BERT, on introduit un **élément plus opaque**. Le score BERT est issu d’un réseau neuronal profond qui transforme la phrase en vecteur de manière non transparente. De plus, il peut *potentiellement* capter des informations corrélées à des caractéristiques sensibles non autorisées (par ex. si le texte mentionne un handicap, un problème de santé, cela pourrait influencer le score). Cette opacité et ce risque de biais implicite sont soulignés dans les recherches récentes : *les LLM comme BERT étant des boîtes noires, il est nécessaire de mettre en place des cadres réglementaires et techniques pour assurer la transparence et éviter les biais discriminatoires*[[48]](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4979155#:~:text=The%20resulting%20BERT,adopting%20advanced%20credit%20risk%20models).

Pour atténuer ce problème, nous avons recours à des méthodes d’**explicabilité locales**. Par exemple, nous avons expérimenté LIME (Local Interpretable Model-agnostic Explanations) sur le modèle BERT : LIME génère des perturbations du texte (enlevant certains mots) et observe l’impact sur la prédiction pour déduire quelles *mots ou phrases du texte ont le plus contribué* au score de risque[[49]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=,Moreover%2C%20our)[[50]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=In%20this%20analysis%2C%20we%20first,top%20fifteen%20words%20derived%20from). Ce type d’analyse a révélé, dans nos tests, que BERT se focalisait sur certains termes clés. Par exemple, la présence de mots comme *“chômage”*, *“en retard de paiement”* avait tendance à augmenter le score de défaut, tandis que des phrases exprimant la stabilité (*“emploi permanent”*, *“revenu fixe”*) le diminuaient. Ce retour explicatif peut être converti en raison compréhensible : *« Le modèle a identifié dans la description du client des éléments suggérant une instabilité financière »* – ce qui, couplé aux autres raisons traditionnelles, peut compléter la fiche explicative pour le décideur ou le client.

Néanmoins, l’explication des modèles NLP reste un défi. Les mots importants identifiés par LIME ou SHAP doivent être interprétés avec prudence et ne remplaceront pas une vraie raison métier tant qu’on n’aura pas validé que le modèle n’utilise pas des proxys indésirables. Dans un cadre opérationnel, il faudrait instaurer une gouvernance autour de l’utilisation de textes : par exemple, filtrer ou anonymiser certaines informations pour éviter d’introduire du biais (ne pas transmettre au modèle des informations comme l’âge, l’origine ethnique, etc. si elles figuraient dans le texte libre, afin de rester conforme à la réglementation).

### Pertinence Opérationnelle et Synthèse Critique

**Apports de la méthode récente** : La mise en œuvre de BERT apporte une **nouvelle source d’information** au processus de décision. Pour Home Credit, cela signifie mieux exploiter les données disponibles sur ces clients à l’historique mince : le texte libre, qui pouvait sembler anecdotique, devient un véritable facteur quantifiable de scoring. Les résultats montrent un potentiel d’amélioration de la détection des risques, en particulier pour certaines niches de prêts. Cela peut se traduire par des décisions plus fines : accepter des clients qui auraient été refusés à tort (faux négatifs réduits) ou refuser des clients risqués qui seraient passés entre les mailles (faux positifs réduits), grâce aux indices textuels. *En finalité, l’approche NLP peut contribuer à la double mission de Home Credit : élargir l’accès au crédit de façon responsable.*

Un avantage non quantifié mais important est la **généralisation** : un modèle comme BERT, entraîné sur un langage très varié, pourrait capter des signaux inattendus sans qu’on les lui spécifie. Par exemple, la tournure de phrase, le niveau de langue, peuvent refléter le profil du client (organisation, sérieux, etc.). Là où un modèle classique nécessiterait de deviner ces indicateurs et de les coder manuellement, BERT les apprend implicitement. Cela rend le modèle potentiellement plus adaptable à des variations de données ou à de nouveaux types de texte, par rapport à une solution figée en un nombre limité de features textuelles.

**Limites et considérations** : D’un autre côté, on doit garder en tête plusieurs limites de l’approche récente avant d’envisager son déploiement opérationnel :

* **Coût et Complexité** : Comme discuté, l’entraînement et la maintenance d’un modèle BERT sont plus lourds. En production, il faudrait surveiller les performances du modèle NLP, éventuellement le re-fine-tuner périodiquement si le langage des demandes évolue. Cela implique de nouvelles compétences au sein de l’équipe (data scientists NLP, MLOps pour déployer du TensorFlow/Transformers en prod). On peut estimer que la complexité est justifiée si le gain en performance l’est aussi, mais pour un faible delta de quelques pourcents d’AUC, chaque organisation devra évaluer le *ROI* d’une telle sophistication.
* **Données disponibles** : Il faut disposer de suffisamment de textes pour entraîner le modèle de manière pertinente. Si seuls 5% des clients renseignent un champ libre, la valeur ajoutée réelle sera limitée, et le modèle BERT risque de n’être applicable qu’à une minorité de cas. Il faudrait dans ce cas encourager la collecte systématique d’une note explicative dans les dossiers (sans rendre cela trop contraignant pour le client). Dans le P2P, les emprunteurs sont volontaires pour fournir ces explications[[5]](https://arxiv.org/html/2401.16458v2#:~:text=An%20inherent%20challenge%20in%20P2P,borrowers%20effectively%20and%20may%20be), mais dans un contexte de crédit à la consommation standard, ce n’est pas toujours la norme.
* **Biais et Fairness** : Un modèle de langage peut involontairement introduire des biais. Par exemple, la manière de s’exprimer peut être corrélée au niveau d’éducation ou à l’origine socioculturelle, qui sont des variables sensibles. Il faudra auditer le modèle pour vérifier qu’il ne discrimine pas indirectement certains groupes de clients. Cela pourrait impliquer de tester le modèle avec des textes simulés identiques à quelques mots près (ex: variation de registre de langue) pour voir si le score change de façon non pertinente. Les auteurs Wu *et al.* (2024) évoquent aussi ce point – la nécessité de cadres transparents vis-à-vis des biais des modèles d’IA en finance[[47]](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4979155#:~:text=the%20BERT,adopting%20advanced%20credit%20risk%20models).
* **Interprétation métier** : Même si l’on parvient à expliquer en partie le fonctionnement de BERT, reste à convaincre les décideurs crédit et les autorités de sa fiabilité. Un modèle black-box peut susciter de la méfiance. Il faudra idéalement présenter des études de cas, des *backtesting* montrant que les décisions prises avec l’aide de BERT auraient évité des pertes ou permis d’accorder plus de bons crédits sans augmenter le risque. Une démarche progressive est conseillée : par exemple, utiliser d’abord le score BERT comme une **information auxiliaire** dans le comité de crédit (sans être automatique), pour voir s’il apporte une valeur ajoutée aux analystes humains, avant d’envisager une automatisation complète.

En définitive, la méthode NLP récente offre une **amélioration mesurée mais réelle** du scoring de crédit pour la population ciblée par Home Credit. Elle semble pertinente pour enrichir l’évaluation des dossiers atypiques ou peu documentés. Cependant, son **adoption opérationnelle** devra s’accompagner de garde-fous en termes d’explicabilité, de biais, et d’infrastructure technique.

## Conclusion

Ce projet de preuve de concept a permis de comparer deux approches de modélisation des données textuelles pour le scoring de crédit Home Credit : une approche classique basée sur un Random Forest et des caractéristiques simplifiées du texte, et une approche moderne exploitant un modèle de langage pré-entraîné (BERT). La démarche s’inscrit dans un contexte où l’innovation est nécessaire pour scorer des clients à l’historique limité en utilisant toutes les données disponibles, y compris les données non structurées.

**Synthèse des résultats** : La méthode NLP récente s’est montrée capable de **mieux comprendre le contenu sémantique** des textes que l’approche manuelle. Elle en tire un score de risque qui a amélioré – légèrement mais significativement – la performance prédictive du modèle de crédit lorsqu’il est combiné aux données classiques. En particulier, l’intégration de BERT a apporté un gain notable pour certains segments de prêts où l’information contextuelle du texte est cruciale. Le Random Forest classique, robuste et éprouvé, reste une référence et continue de fournir l’essentiel du pouvoir prédictif grâce aux variables structurelles, mais il a bénéficié de l’apport du texte analysé par BERT.

**Avantages** : La méthode récente permet d’exploiter finement une source de données jusqu’alors sous-utilisée, renforçant ainsi la capacité du modèle à différencier des profils risqués de profils sains parmi les clients sans historique. Elle peut aider Home Credit à accorder des crédits à de “bons” clients autrefois refusés par manque d’information, améliorant l’inclusion financière tout en maintenant le risque sous contrôle. C’est donc un atout potentiellement différenciant dans un marché du crédit concurrentiel, en particulier pour adresser la clientèle des nouveaux emprunteurs.

**Limites** : D’un autre côté, cette sophistication entraîne une complexité accrue et pose des défis en termes d’explicabilité. L’implémentation devra être accompagnée de mesures garantissant la transparence du modèle (audit régulier des décisions, outils d’explication locaux type LIME/SHAP) et le respect des contraintes réglementaires (pas d’utilisation indue d’informations biaisées). Il faudra également veiller à la stabilité du modèle NLP dans le temps – les modèles de langage peuvent dériver si le type de langage employé par les clients évolue, ou si des expressions nouvelles apparaissent. Une veille technologique et une actualisation périodique (par exemple, re-entraîner ou affiner le modèle chaque année avec de nouveaux exemples) feront partie des bonnes pratiques. Notons qu’avec l’émergence rapide des **LLM encore plus récents** (ex: GPT-3, GPT-4) on peut anticiper que des modèles de langue spécialisés pour la finance (“FinBERT”, etc.) deviendront de plus en plus accessibles, ce qui pourra encore améliorer la performance ou faciliter l’explicabilité (certains modèles étant entraînés à justifier leurs prédictions).

**Recommandations opérationnelles** : Il est recommandé d’envisager une **phase pilote** pour introduire le score BERT dans le processus Home Credit. Par exemple, utiliser le score comme une variable auxiliaire dans un outil interne pendant quelques mois pour accumuler du feedback. On pourra mesurer si les prêts accordés/refusés diffèrent sensiblement avec ce score, et si les taux de défaut évoluent favorablement. Parallèlement, il faudra former les analystes aux bases de cette nouvelle approche pour qu’ils intègrent son usage en toute confiance.

En conclusion, l’approche NLP récente constitue un progrès méthodologique intéressant pour le scoring de crédit dans des contextes d’**information limitée**. Notre étude de cas montre que, tout en s’appuyant sur les fondations solides d’un Random Forest classique, l’apport d’un modèle de langage pré-entraîné comme BERT enrichit la modélisation et ouvre de nouvelles perspectives pour affiner l’évaluation du risque client. Utilisée de manière réfléchie et encadrée, cette méthode pourrait accroître la **pertinence opérationnelle** du scoring Home Credit en améliorant sa **discrimination** et son **inclusivité**, permettant ainsi de mieux servir la clientèle tout en maîtrisant le risque.

## Bibliographie

* **Wu, Z.**, **Dong, Y.**, **Li, Y.**, & **Shi, B.** (2024). *Unleashing the power of text for credit default prediction: Comparing human-generated and AI-generated texts*. (Working Paper, March 2024)[[4]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=of%20big%20data%2C%20several%20recent,modelling%20process%20can%20significantly%20improve)[[50]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=In%20this%20analysis%2C%20we%20first,top%20fifteen%20words%20derived%20from).
* **Sanz-Guerrero, M.** & **Arroyo, J.** (2024). *Credit Risk Meets Large Language Models: Building a Risk Indicator from Loan Descriptions in P2P Lending*. arXiv preprint arXiv:2401.16458[[27]](https://arxiv.org/html/2401.16458v2#:~:text=Peer,to%20analyze%20borrowers%E2%80%99%20loan%20descriptions)[[36]](https://arxiv.org/html/2401.16458v2#:~:text=Table%2013%20%20also%20shows,information%20for%20the%20classification%20task).
* **Narayanan, D.** (2021). *Home Credit Default Risk (Part 1) : Business Understanding, Data Cleaning and EDA*. Analytics Vidhya/Medium, Jul. 25, 2021[[2]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=Obviously%2C%20this%20is%20a%20massive,unreasonably%20high%20rates%20of%20interest)[[3]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=In%20order%20to%20address%20this,not%20have%20their%20applications%20rejected).
* **Devlin, J.**, **Chang, M.-W.**, **Lee, K.**, & **Toutanova, K.** (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of NAACL 2019. (arXiv:1810.04805)
* **Alammar, J.** (2019). *The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)*[[23]](https://jalammar.github.io/illustrated-bert/#:~:text=Each%20position%20outputs%20a%20vector,CLS%5D%20token%20to) (Blog post: jalammar.github.io/illustrated-bert/).

[[1]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=%E2%80%98Home%20Credit%E2%80%99%20Group%20is%20a,online%20as%20well%20as%20offline) [[2]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=Obviously%2C%20this%20is%20a%20massive,unreasonably%20high%20rates%20of%20interest) [[3]](https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c#:~:text=In%20order%20to%20address%20this,not%20have%20their%20applications%20rejected) Home Credit Default Risk (Part 1) : Business Understanding, Data Cleaning and EDA | by Dhruv Narayanan | Analytics Vidhya | Medium

<https://medium.com/analytics-vidhya/home-credit-default-risk-part-1-business-understanding-data-cleaning-and-eda-1203913e979c>

[[4]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=of%20big%20data%2C%20several%20recent,modelling%20process%20can%20significantly%20improve) [[12]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=underwriting%2C%20and%20greater%20financial%20inclusion,leverage%20text%20data%2C%20a%20major) [[49]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=,Moreover%2C%20our) [[50]](https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf#:~:text=In%20this%20analysis%2C%20we%20first,top%20fifteen%20words%20derived%20from) crc.business-school.ed.ac.uk

<https://www.crc.business-school.ed.ac.uk/sites/crc/files/2024-06/Unleashing-the-power-of-text-for-credit-default-prediction_1.pdf>

[[5]](https://arxiv.org/html/2401.16458v2#:~:text=An%20inherent%20challenge%20in%20P2P,borrowers%20effectively%20and%20may%20be) [[6]](https://arxiv.org/html/2401.16458v2#:~:text=lenders,20) [[7]](https://arxiv.org/html/2401.16458v2#:~:text=Traditional%20credit%20scoring%20models%20do,7) [[8]](https://arxiv.org/html/2401.16458v2#:~:text=in%20the%20narratives%20submitted%20by,7) [[9]](https://arxiv.org/html/2401.16458v2#:~:text=measure%20aspects%20such%20as%20polarity%2C,correlate%20positively%20with%20loan%20repayment) [[10]](https://arxiv.org/html/2401.16458v2#:~:text=logistic%20regression,correlate%20positively%20with%20loan%20repayment) [[11]](https://arxiv.org/html/2401.16458v2#:~:text=both%20) [[13]](https://arxiv.org/html/2401.16458v2#:~:text=indicator%20variables,correlate%20positively%20with%20loan%20repayment) [[14]](https://arxiv.org/html/2401.16458v2#:~:text=,improved%20classification%20in%20other%20fields) [[15]](https://arxiv.org/html/2401.16458v2#:~:text=We%20will%20focus%20on%20BERT,3%20model%20has%20175) [[17]](https://arxiv.org/html/2401.16458v2#:~:text=) [[18]](https://arxiv.org/html/2401.16458v2#:~:text=which%20is%20a%20Transformer,3%20model%20has%20175) [[22]](https://arxiv.org/html/2401.16458v2#:~:text=To%20further%20elucidate%20the%20role,tuned%20for%20biomedical) [[25]](https://arxiv.org/html/2401.16458v2#:~:text=linguistic%20model%20for%20various%20of,suited%20for%20language%20generation%20tasks) [[27]](https://arxiv.org/html/2401.16458v2#:~:text=Peer,to%20analyze%20borrowers%E2%80%99%20loan%20descriptions) [[28]](https://arxiv.org/html/2401.16458v2#:~:text=We%20apply%20transfer%20learning%20to,refining%20credit%20risk%20assessment%20methodologies) [[29]](https://arxiv.org/html/2401.16458v2#:~:text=resulting%20BERT,refining%20credit%20risk%20assessment%20methodologies) [[30]](https://arxiv.org/html/2401.16458v2#:~:text=lenders,20) [[31]](https://arxiv.org/html/2401.16458v2#:~:text=However%2C%20these%20methods%20have%20limitations,and%20processing%20complex%20textual%20data) [[34]](https://arxiv.org/html/2401.16458v2#:~:text=match%20at%20L850%20threshold%20to,of%20the%20instances) [[35]](https://arxiv.org/html/2401.16458v2#:~:text=threshold%20to%20binarize%20the%20BERT,of%20the%20instances) [[36]](https://arxiv.org/html/2401.16458v2#:~:text=Table%2013%20%20also%20shows,information%20for%20the%20classification%20task) [[37]](https://arxiv.org/html/2401.16458v2#:~:text=the%20BERT%20score%20in%20terms,information%20for%20the%20classification%20task) [[38]](https://arxiv.org/html/2401.16458v2#:~:text=Metric%20Quant,5724) [[39]](https://arxiv.org/html/2401.16458v2#:~:text=Figure%207%20illustrates%20the%20changes,more%20precise%20delineation%20of%20the) [[40]](https://arxiv.org/html/2401.16458v2#:~:text=Figure%207%20illustrates%20the%20changes,more%20precise%20delineation%20of%20the) [[41]](https://arxiv.org/html/2401.16458v2#:~:text=default%20risk,obtaining%20funds%20for%20a%20business) [[42]](https://arxiv.org/html/2401.16458v2#:~:text=7) [[43]](https://arxiv.org/html/2401.16458v2#:~:text=It%20is%20noteworthy%20that%20Figure,variables%20influence%20the%20target%20class) [[44]](https://arxiv.org/html/2401.16458v2#:~:text=Figure%206%20reveals%20a%20direct,as%20strong%20indicators%20of%20default) [[45]](https://arxiv.org/html/2401.16458v2#:~:text=threshold%20to%20binarize%20the%20BERT,of%20the%20instances) Credit Risk Meets Large Language Models: Building a Risk Indicator from Loan Descriptions in P2P Lending

<https://arxiv.org/html/2401.16458v2>

[[16]](https://en.wikipedia.org/wiki/BERT_(language_model)#:~:text=learning%20,3) [[19]](https://en.wikipedia.org/wiki/BERT_(language_model)#:~:text=BERT%20is%20trained%20by%20masked,3) [[20]](https://en.wikipedia.org/wiki/BERT_(language_model)#:~:text=BERT%20was%20originally%20implemented%20in,7) BERT (language model) - Wikipedia

<https://en.wikipedia.org/wiki/BERT_(language_model)>

[[21]](https://www.projectpro.io/article/bert-nlp-model-explained/558#:~:text=For%20example%2C%20let%E2%80%99s%20take%20the,following%20words) BERT NLP Model Explained for Complete Beginners

<https://www.projectpro.io/article/bert-nlp-model-explained/558>

[[23]](https://jalammar.github.io/illustrated-bert/#:~:text=Each%20position%20outputs%20a%20vector,CLS%5D%20token%20to) [[24]](https://jalammar.github.io/illustrated-bert/#:~:text=Image) The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) – Jay Alammar – Visualizing machine learning one concept at a time.

<https://jalammar.github.io/illustrated-bert/>

[[26]](https://openreview.net/attachment?id=X3irT6Yua8&name=pdf#:~:text=The%20fine,Analysis%20of%20the%20BERT%20score) [[32]](https://openreview.net/attachment?id=X3irT6Yua8&name=pdf#:~:text=As%20outlined%20in%20Section%203%2C,to%20enhance%20the%20model%E2%80%99s%20adaptabil%02ity) [[33]](https://openreview.net/attachment?id=X3irT6Yua8&name=pdf#:~:text=knowledge,for%20all%20the%20dropout%20layers) Credit Risk Meets Large Language Models: Building a Risk Indicator from Loan Descriptions in P2P Lending

<https://openreview.net/attachment?id=X3irT6Yua8&name=pdf>

[[46]](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4979155#:~:text=descriptions,of%20trust%20and%20transparency%20in) [[47]](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4979155#:~:text=the%20BERT,adopting%20advanced%20credit%20risk%20models) [[48]](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4979155#:~:text=The%20resulting%20BERT,adopting%20advanced%20credit%20risk%20models) Credit Risk Meets Large Language Models: Building a Risk Indicator from Loan Descriptions in Peer-to-Peer Lending by Mario Sanz-Guerrero, Javier Arroyo :: SSRN

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4979155>