EPITA
École d'Ingénieurs pour l'Informatique et les Techniques Avancées

Majeure SCIA-G
Sciences Cognitives, Intelligence Artificielle & Graphes

Projet Prédiction Conforme

Implémentation et Analyse dans le Domaine Financier

RAPPORT DE PROJET

Auteur

Erwann Lesech, Étudiant

Encadrant

Rémi Vaucher, PhD — Professeur

Lyon, le 12 novembre 2025

Présenté dans le cadre du cours *Stochastiques* — Promo 2026

Contents

1 Introduction **2**

2 Tâche de régression - Risque d'un client contracteur de crédit **2**
2.1 Présentation du jeu de données et pré-traitement 2

3 Régression Quantile **7**
3.1 Fondements théoriques 7
3.2 Choix du modèle et justification 7
3.3 Mise en place et construction des intervalles 7
3.4 Résultats et interprétation 7

4 Prédiction Conforme pour la Régression **9**
4.1 Choix des méthodes et justification 9
4.2 Mise en place des modèles 10
4.3 Résultats et interprétation 11

5 Tâche de classification - Catégorie de note de crédit d'entreprise **13**
5.1 Présentation du jeu de données et pré-traitement 13

6 Prédiction Conforme pour la Classification **18**
6.1 Choix des méthodes et justification 18
6.2 Mise en place des modèles 18
6.3 Résultats et interprétation 19

7 Conclusion **21**

1 Introduction

Contexte et enjeux de la quantification d'incertitude

Historiquement, la **régression quantile** a constitué une première approche pour estimer des intervalles statistiques autour d'une prédiction. Elle permettait déjà de quantifier l'incertitude associée à une estimation en fournissant des bornes de confiance sur la variable cible. Avec l'émergence du machine learning, les modèles se sont complexifiés et leurs performances ont considérablement augmenté, mais la question de la fiabilité des prédictions est restée centrale.

Les **modèles d'apprentissage automatique** se concentrent souvent sur la recherche de la meilleure précision possible, sans nécessairement indiquer dans quelle mesure leurs prédictions peuvent être considérées comme fiables. Or, dans des domaines critiques comme la santé, la finance ou la cybersécurité, une erreur de prédiction peut avoir des conséquences majeures. Il devient donc essentiel de mesurer non seulement la valeur prédite, mais aussi le degré d'incertitude qui l'accompagne.

C'est alors que la **prédiction conforme** répond précisément à ce besoin. Elle fournit un apport mathématique pour associer à chaque prédiction une estimation de confiance, tout en garantissant statistiquement un taux d'erreur contrôlé.

Problématique du secteur financier

Le secteur financier est un exemple typique où la prise de décision repose sur des modèles prédictifs sensibles à l'incertitude. Les prévisions de risque, les notations de crédit ou l'évaluation des performances d'entreprises reposent sur des données volatiles et complexes, souvent influencées par des facteurs externes difficiles à anticiper. Dans un tel contexte, un modèle classique fournissant une seule prédiction numérique ou catégorielle peut s'avérer insuffisant, voire trompeur.

L'intégration de la prédiction conforme dans ce domaine permet de mieux maîtriser la confiance associée aux estimations, qu'il s'agisse d'évaluer **la solvabilité d'une entreprise** ou **le risque financier d'un client**.

La suite de ce rapport s'appuie sur deux ensembles de données issus du domaine financier, permettant d'illustrer l'apport concret de la prédiction conforme dans des contextes réels de classification et de régression.

2 Tâche de régression - Risque d'un client contracteur de crédit

2.1 Présentation du jeu de données et pré-traitement

Présentation et choix du jeu de données

Le jeu de données *Financial Risk for Loan Approval*[1] contient 20 000 observations de demandes de crédit avec 36 variables décrivant les caractéristiques financières et personnelles des emprunteurs. La variable cible `RiskScore` représente un score de risque continu (28.8 à 84.0), rendant ce dataset particulièrement adapté aux techniques de régression quantile et de prédiction conforme.

Ce dataset présente plusieurs avantages pour notre étude. Avec plus de 10 000 observations disponibles pour l'entraînement après un split train/test classique, nous disposons d'un ensemble de calibration suffisamment large pour garantir la validité statistique de la prédiction

[1] `https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval`

conforme. La taille du dataset assure également une estimation robuste des différents quantiles en régression quantile. Par ailleurs, l'absence de structure temporelle dans les données et l'indépendance entre les observations (chaque demande de crédit correspond à un client distinct) satisfont l'hypothèse d'échangeabilité nécessaire pour les garanties théoriques de la prédiction conforme. Enfin, le caractère synthétique du dataset élimine les contraintes de confidentialité tout en préservant des distributions réalistes de variables financières, permettant une analyse approfondie sans limitation d'accès aux données.

Description détaillée des données

Le dataset comprend 20 000 observations réparties sur 36 variables, sans aucune valeur manquante. Les variables se décomposent en 29 variables numériques, 6 variables catégorielles et 1 variable temporelle (`ApplicationDate`) qui sera exclue de l'analyse.

La variable cible `RiskScore` est une variable continue représentant le niveau de risque associé à chaque demande de crédit. Elle s'étend de 28.8 à 84.0 avec une moyenne de 50.8 et un écart-type de 7.8. La distribution présente une légère asymétrie à droite, avec des quartiles à Q1=46, Q2=52 et Q3=56, suggérant une concentration des scores autour de la médiane avec quelques valeurs extrêmes vers les risques élevés.

Les 35 variables explicatives couvrent l'ensemble des dimensions pertinentes pour l'évaluation du risque de crédit, comme synthétisé dans le tableau ci-dessous.

Catégorie	**Type**	**Variables principales**
Profil démographique	Numérique	`Age`, `Experience`, `NumberOfDependents`, `JobTenure`
	Catégorielle	`EducationLevel`, `EmploymentStatus`, `MaritalStatus`
Revenus & Patrimoine	Numérique	`AnnualIncome`, `MonthlyIncome`, `TotalAssets`, `NetWorth`, `SavingsAccountBalance`, `CheckingAccountBalance`
Historique de crédit	Numérique	`CreditScore`, `LengthOfCreditHistory`, `PaymentHistory`, `PreviousLoanDefaults`, `BankruptcyHistory`
Crédit en cours	Numérique	`NumberOfOpenCreditLines`, `NumberOfCreditInquiries`, `CreditCardUtilizationRate`
Dette	Numérique	`TotalLiabilities`, `MonthlyDebtPayments`, `DebtToIncomeRatio`, `TotalDebtToIncomeRatio`
Demande de prêt	Numérique	`LoanAmount`, `LoanDuration`, `MonthlyLoanPayment`, `InterestRate`, `BaseInterestRate`
	Catégorielle	`LoanPurpose`, `HomeOwnershipStatus`
Autres	Numérique	`UtilityBillsPaymentHistory`

Table 1: Synthèse des variables du dataset de régression

La variable `LoanApproved`, présente dans le dataset original, représente une décision binaire d'approbation du prêt. Cette variable étant un label alternatif (classification), elle est exclue de notre analyse de régression pour éviter toute fuite d'information et se concentrer uniquement sur la prédiction du score de risque continu.

Problématique métier et justification

Considérons le scénario d'un conseiller bancaire face à une demande de crédit. Un modèle de machine learning classique lui fournirait une prédiction unique : « Ce client a un score de risque estimé à 48 ». Cette information, bien qu'utile, reste insuffisante pour prendre une décision éclairée. Le conseiller ne dispose d'aucune indication sur la fiabilité de cette prédiction. S'agit-il d'un profil typique, bien compris par le modèle, ou d'un cas atypique où l'incertitude est élevée ?

La **régression quantile** enrichit cette analyse en produisant de intervalles. Si l'intervalle est étroit, le profil est bien caractérisé et une décision automatisée peut être envisagée. Si l'intervalle est large, cela signale une forte variabilité du risque et justifie un examen approfondi. Cette information permet également d'adapter les conditions du prêt : un client au quantile élevé pourrait se voir proposer un taux d'intérêt majoré ou des garanties supplémentaires.

La **prédiction conforme** va plus loin en fournissant une garantie statistique : « Avec 90% de confiance, le véritable score de risque se situe dans l'intervalle [45, 51] ». Contrairement à un modèle classique qui ne contrôle pas son taux d'erreur, la prédiction conforme garantit que 90% des clients auront leur score réel couvert par l'intervalle prédit. Cette propriété est cruciale dans un contexte réglementaire où les institutions financières doivent justifier leurs décisions et maîtriser leur exposition au risque. Un intervalle trop large pour un client donné peut déclencher une analyse manuelle, tandis qu'un intervalle étroit valide la confiance du modèle et accélère le processus d'approbation.

Analyse exploratoire des données (EDA)

L'analyse exploratoire révèle plusieurs caractéristiques importantes du dataset. La distribution de la variable cible `RiskScore` présente une légère asymétrie à droite avec quelques valeurs aberrantes au-delà de 70, suggérant l'existence de profils à très haut risque peu fréquents mais à surveiller (Figure 1).

Figure 1: Distribution de la variable cible RiskScore (histogramme et boxplot)

L'analyse de corrélation met en évidence plusieurs variables fortement liées entre elles. Les variables `MonthlyIncome` et `AnnualIncome` affichent logiquement une corrélation très élevée (0.99), tout comme `Age` et `Experience` (0.983). L'analyse bivariée (Figure 3) révèle une variable visuellement très corrélée à la cible : `MonthlyIncome`. Cela semble logique côté métier, le premier critère d'évaluation du risque étant la capacité de remboursement liée aux revenus.

Nous risquons donc d'utiliser cette variable comme exemple pour notre régression quantile plus tard.

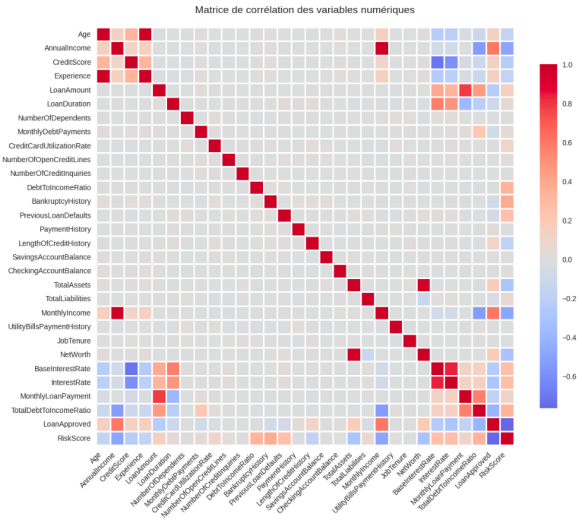

Figure 2: Matrice de corrélation des variables numériques

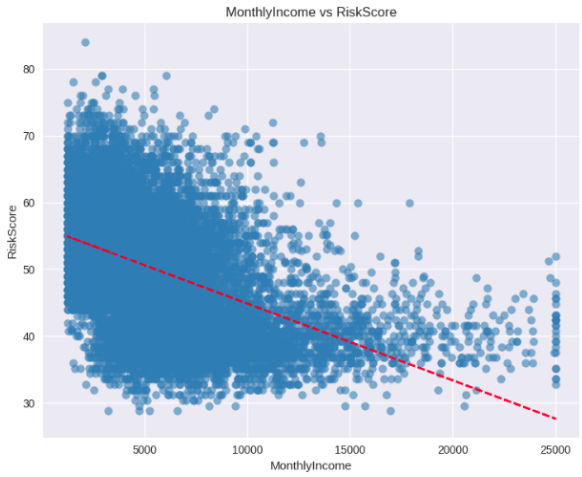

Figure 3: Analyse bivariée : relations entre features principales et RiskScore

Enfin, le graphique des corrélations avec la cible (Figure 4) permet d'identifier les variables les plus prédictives. Hormis `LoanApproved` qui est exclue de l'analyse, on retrouve à nouveau la variable `MonthlyIncome` (-0.487), ainsi que `AnnualIncome` (-0.483) et `BankruptcyHistory` (0.378) comme les features les plus corrélées à `RiskScore`.

Figure 4: Corrélations des features avec la variable cible RiskScore

Préparation et prétraitement des données

Le prétraitement des données s'effectue en plusieurs étapes successives. La variable `ApplicationDate`, représentant une série temporelle synthétique sans valeur prédictive, est d'abord supprimée. De même, la variable `LoanApproved`, qui correspond à un label binaire d'approbation de prêt et constitue donc une tâche de classification alternative, est exclue de l'analyse pour se concentrer uniquement sur la prédiction du score de risque continu.

Les variables catégorielles (`EmploymentStatus`, `EducationLevel`, `MaritalStatus`, `LoanPurpose`, `HomeOwnershipStatus`) sont ensuite encodées en utilisant un encodage one-hot, créant des variables binaires pour chaque modalité tout en supprimant une modalité de référence pour éviter la multicolinéarité. Les variables numériques sont normalisées par standardisation (moyenne 0, écart-type 1) afin d'homogénéiser les échelles et d'améliorer la convergence des algorithmes

d’apprentissage.

3 Régression Quantile

3.1 Fondements théoriques

La régression quantile modélise les quantiles conditionnels de la variable cible plutôt que sa moyenne. Pour un quantile $\tau \in (0, 1)$, elle résout :

$$\hat{\beta}_\tau = \arg\min_\beta \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta)$$

où $\rho_\tau(u)$ est la fonction de perte quantile (pinball loss) définie par :

$$\rho_\tau(u) = u(\tau - \mathbf{1}_{u<0}) = \begin{cases} \tau u & \text{si } u \geq 0 \\ (\tau - 1)u & \text{si } u < 0 \end{cases}$$

3.2 Choix du modèle et justification

Nous avons implémenté une régression quantile linéaire via `QuantileRegressor` de scikit-learn.

Trois configurations ont été testées : (1) une seule feature (`MonthlyIncome`), permettant une visualisation 2D, (2) toutes les features disponibles (33 après preprocessing), (3) les features fortement corrélées ($|\text{corr}| \geq 0.3$). Une quatrième configuration ajoute une **extension polynomiale de degré 2** au modèle 1 pour vérifier l'existence de non-linéarités quadratiques entre le revenu et le risque.

3.3 Mise en place et construction des intervalles

Le prétraitement suit le protocole établi : encodage one-hot des catégorielles et standardisation des numériques. Le split train/test (67%/33%) garantit une évaluation robuste sur 6 600 observations.

Pour un niveau de confiance $1 - \beta = 90\%$ ($\beta = 0.1$), trois régressions quantiles sont entraînées par configuration : $Q_{0.05}$ (borne inférieure), $Q_{0.50}$ (médiane conditionnelle) et $Q_{0.95}$ (borne supérieure). L'intervalle de prédiction est construit comme $[Q_{0.05}(X_{new}), Q_{0.95}(X_{new})]$, devant théoriquement contenir 90% des observations.

3.4 Résultats et interprétation

Métriques de performance & résultats

Le **taux de couverture empirique** mesure la calibration du modèle :

$$\text{Coverage} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbb{1}_{Q_{0.05}(X_i) \leq y_i \leq Q_{0.95}(X_i)}$$

Un modèle bien calibré doit afficher un taux proche de 90%. Une procédure de bootstrap de 50 itérations permet d'assurer la stabilité de cette mesure au dépend de l'aléatoire.

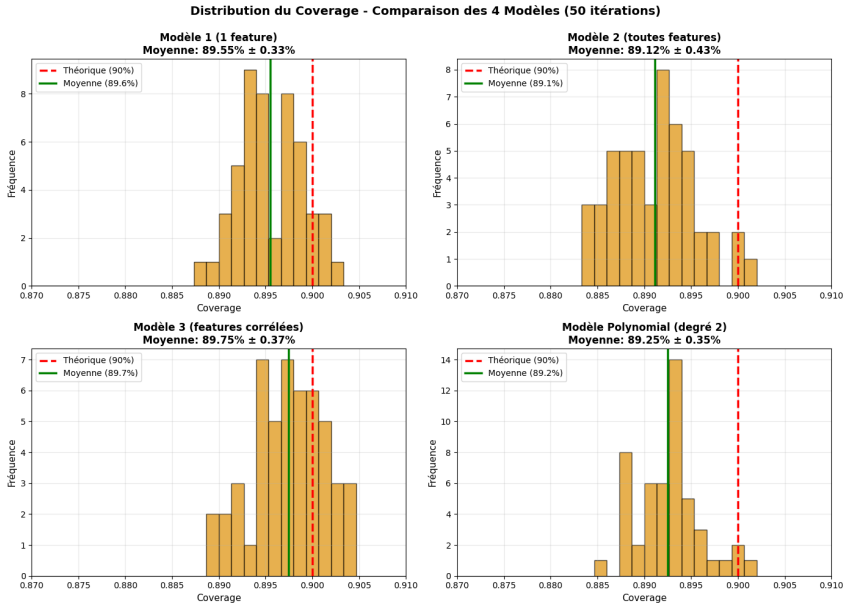

Figure 5: Taux de couverture empirique des quatre modèles de régression quantile

D'autres métriques comme la `RMSE` ou la `MAE` ont été utilisé pour comparer la précision des modèles (basé seulement sur le quantile médian). Enfin la largeur moyenne des intervalles ont aussi été mesurés comme montré ci-contre:

Modèle	**MAE**	**RMSE**	**Coverage (%)**	**Largeur moy.**
Modèle 1 (1 feature linéaire)	5.29	6.84	89.44 ± 0.37	22.32
Modèle 2 (toutes features)	**2.77**	**3.88**	89.12 ± 0.36	**10.89**
Modèle 3 (features corrélées)	3.84	5.14	$\mathbf{89.82 \pm 0.37}$	16.38
Modèle 4 (1 feature polynomiale)	5.23	6.75	89.26 ± 0.35	22.25

Table 2: Comparaison des performances des modèles de régression quantile

Analyse comparative

Le **modèle 2** (toutes features) offre la meilleure précision prédictive (MAE=2.77, RMSE=3.88) avec les intervalles les plus étroits (10.89 points), confirmant l'apport d'une information multivariée complète. Sa couverture légèrement inférieure (89.12%) suggère néanmoins des difficultés sur certains profils atypiques.

Le **modèle 3** (features corrélées) constitue le meilleur compromis : couverture optimale (89.82%), précision intermédiaire (MAE=3.84) et largeur raisonnable (16.38 points). Il démontre qu'une sélection rigoureuse de variables ($|\text{corr}| \geq 0.3$) suffit à capturer l'essentiel de l'information prédictive.

Les **modèles 1 et 4** (linéaire vs polynomial) présentent des performances quasi-identiques. L'amélioration marginale du modèle polynomial (MAE 5.23 vs 5.29) valide que la relation `MonthlyIncome–RiskScore` est **essentiellement linéaire**, avec peu de non-linéarité quadratique à exploiter.

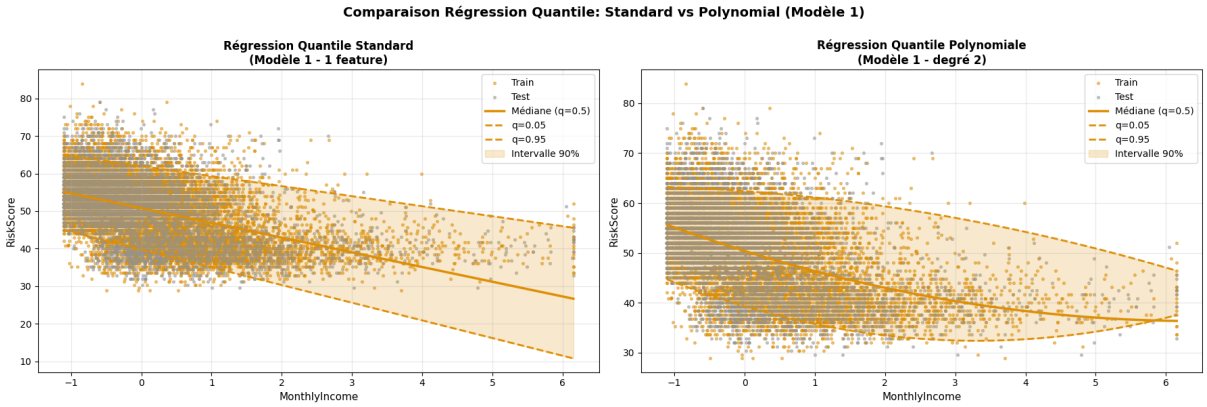

Figure 6: Régression quantile sur le modèle 1 (une seule feature) : quantiles 0.05, 0.50 et 0.95

Tous les modèles affichent une couverture proche de 90% avec une faible variabilité ($\pm$ 0.35-0.37%), démontrant la fiabilité de la régression quantile pour construire des intervalles bien calibrés.

Plus-value métier

La régression quantile fournit une information cruciale pour la décision de crédit en produisant trois scénarios : (1) **optimiste** ($Q_{0.05}$) pour le pricing agressif, (2) **central** ($Q_{0.50}$) pour la décision standard, (3) **pessimiste** ($Q_{0.95}$) pour l'évaluation du risque maximal.

Un conseiller bancaire peut adapter les conditions du prêt selon la largeur de l'intervalle. Un intervalle étroit signale une prédiction fiable autorisant une décision automatisée, tandis qu'un intervalle large indique une forte incertitude justifiant un examen manuel approfondi. Le choix du modèle dépend du contexte opérationnel : modèle 1 pour l'interprétabilité maximale et la communication client, modèle 3 pour l'équilibre précision/simplicité, modèle 2 pour la performance maximale en production.

4 Prédiction Conforme pour la Régression

4.1 Choix des méthodes et justification

Nous avons implémenté deux variantes de prédiction conforme pour la régression, chacune adaptée à des contextes différents.

Split Conformal Prediction (SCP)

SCP est la méthode la plus simple et directe. Elle divise les données en trois ensembles disjoints : train (50%), calibration (40%) et test (10%). Le modèle est entraîné une fois sur le train set, puis calibré sur le calibration set pour calculer le quantile $\hat{q}$ des scores de non-conformité. Cette approche est efficace avec des datasets volumineux (> 10 000 obs.) car le split ne réduit pas significativement la taille effective. Son avantage est sa rapidité (un seul entraînement) et sa simplicité d'implémentation.

Cross-Validation Plus (CV+)

CV+ exploite mieux les données via validation croisée K-folds. Elle entraîne K modèles (ici $K = 3$) sur des folds différents et collecte les scores de non-conformité sur les prédictions out-of-sample de chaque fold. Le quantile est calculé sur l'ensemble agrégé de ces scores. Pour la prédiction finale, les K modèles sont moyennés, réduisant la variance. Cette approche est plus robuste sur datasets moyens (2 000–10 000 obs.) en exploitant mieux les données disponibles, au prix d'un coût K fois supérieur.

Pourquoi pas Jackknife+ et Full Conformal Prediction ?

Jackknife+ nécessite n entraînements en leave-one-out (20 000 sur notre dataset), soit un coût prohibitif sans gain pratique significatif.

Full Conformal Prediction (FCP) requiert de tester toutes les valeurs candidates y_{new} en réentraînant le modèle à chaque fois, rendant la méthode inapplicable en production.

SCP et **CV+** offrent le meilleur compromis rigueur/applicabilité pour l'évaluation du risque de crédit.

4.2 Mise en place des modèles

Division des données et prétraitement

Le preprocessing suit le protocole standard : encodage one-hot des catégorielles et standardisation des numériques. Pour SCP, split en trois ensembles disjoints : train (50%, 10 000 obs.), calibration (40%, 8 000 obs.) et test (10%, 2 000 obs.). Pour CV+, fusion train+calibration (90%, 18 000 obs.) utilisée pour validation croisée 3-folds, avec test set identique pour comparaison équitable.

Choix de l'algorithme ML de base

Nous avons utilisé une **régression Ridge** (`Ridge`, $\alpha = 1.0$) comme modèle de base $\hat{f}$. Ridge offre stabilité via régularisation L2, évitant le surapprentissage tout en restant interprétable. Sa rapidité d'entraînement est cruciale pour CV+. Des modèles plus complexes (`RandomForest`, `GradientBoosting`) auraient de meilleures performances mais un temps d'entraînement trop long pour cette étude.

Principe commun : score de conformité

Les deux méthodes partagent la même logique fondamentale. Le **score de conformité** mesure l'écart entre prédiction et valeur réelle :

$$S(X_i, y_i) = |y_i - \hat{f}(X_i)|$$

Nous utilisons l'erreur absolue (L1), symétrique et robuste aux outliers. Sur l'ensemble de calibration (ou via CV), on calcule les scores $S_1, \ldots, S_n$ et leur quantile ajusté :

$$\hat{q} = \text{Quantile}\left(S_1, \ldots, S_n; \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right)$$

L'intervalle de prédiction pour X_{new} est :

$$\hat{C}(X_{new}) = [\hat{f}(X_{new}) - \hat{q},\ \hat{f}(X_{new}) + \hat{q}]$$

4.3 Résultats et interprétation

Synthèse des performances

Sur le test set (2 000 observations), les deux méthodes respectent la garantie théorique de 90% de couverture :

Méthode	Couverture	Largeur moy.	MAE	R^2
SCP	89.15%	11.85	3.94	0.785
CV+	**90.20%**	**6.16**	**2.89**	**0.940**

Table 3: Comparaison SCP vs CV+ pour la prédiction conforme en régression

Figure 7: Exemples de prédictions avec intervalles conformes pour SCP et CV+ sur le test set

CV+ surpasse SCP sur tous les critères : couverture optimale (90.20% vs 89.15%), intervalles 48% plus étroits (6.16 vs 11.85), et meilleures performances prédictives (R^2=0.940 vs 0.785). L'exploitation de la validation croisée permet à CV+ d'utiliser 18 000 observations au lieu de 10 000 pour SCP, expliquant ces améliorations. Pourtant, la SCP présente déjà des résultats probants.

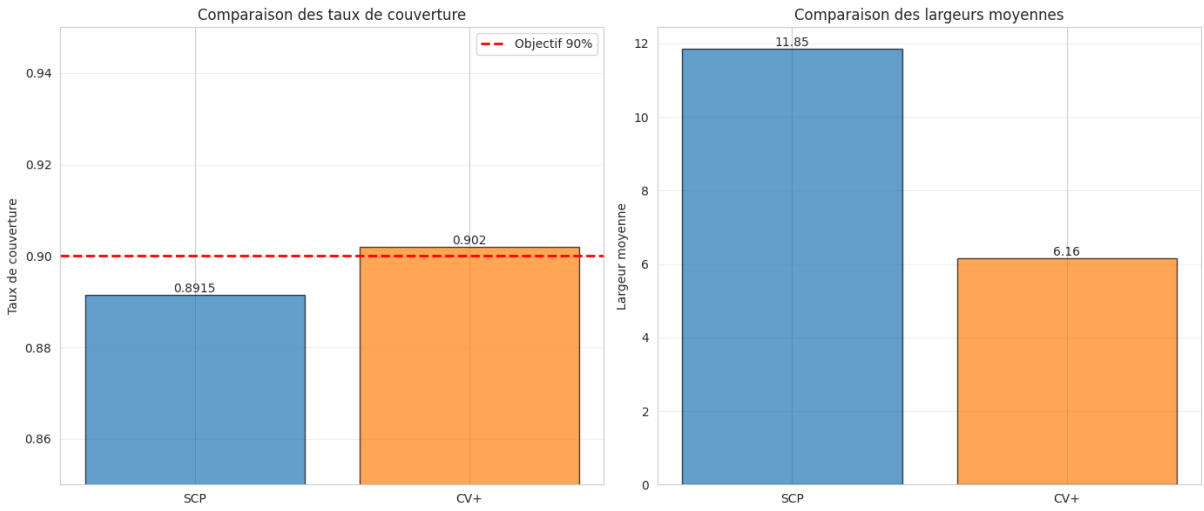

Figure 8: Comparaison des taux de couverture et largeurs moyennes d'intervalles pour SCP et CV+

Intérêt métier

À l'instar de la régression quantile, la prédiction conforme fournit non pas une valeur ponctuelle mais un intervalle prédictif contenant la valeur réelle avec une probabilité garantie de 90%. Pour une banque, cela se traduit par :

- décision automatisée si le score et son intervalle restent en-deçà d'un seuil de risque interne ;
- analyse approfondie si l'intervalle est large ou chevauche des zones à risque ;
- rejet si l'intervalle dépasse les niveaux tolérés.

De plus, la largeur et la position de l'intervalle permettent d'ajuster les *conditions de prêt* (taux, garanties, plafond) : un intervalle favorable peut améliorer l'offre, un intervalle défavorable ou incertain justifie des conditions plus strictes.

Limites et perspectives

Les intervalles construits ici sont symétriques et de largeur constante, ce qui ne capte pas l'hétérogénéité de l'incertitude par profil. Une amélioration est l'utilisation de la **Conformal Quantile Regression (CQR)** pour obtenir des intervalles adaptatifs et asymétriques tout en conservant des garanties conformes.

Dans notre cas d'étude, l'implémentation de CQR a été tentée mais freinée par une capacité mémoire insuffisante : la réduction du volume de données pour l'entraînement a dégradé la qualité des résultats. Il conviendrait donc de réessayer CQR avec une machine plus performante pour valider son bénéfice.

5 Tâche de classification - Catégorie de note de crédit d'entreprise

5.1 Présentation du jeu de données et pré-traitement

Présentation et choix du jeu de données

Le jeu de données *Corporate Credit Rating*[2] contient 2 031 observations de notations de crédit d'entreprises américaines avec 31 variables décrivant leurs performances financières et opérationnelles. La variable cible `Rating` représente une notation de crédit catégorielle (10 classes : AAA, AA, A, BBB, BB, B, CCC, CC, C, D), rendant ce dataset particulièrement adapté aux techniques de prédiction conforme pour la classification.

Ce dataset présente plusieurs avantages pour notre étude. Bien que plus petit que le dataset de régression, il offre suffisamment d'observations pour garantir la validité statistique de la prédiction conforme après calibration (> 1000 données) et peut nous permettre éventuellement d'essayer d'appliquer des algorithmes permettant un ensemble de calibration réduit. La présence de notations issues de quatre agences de notation réputées (Standard & Poor's, Moody's, Fitch Ratings, Egan-Jones) assure la crédibilité et la cohérence des labels. L'absence de structure temporelle dans les données et l'indépendance entre les observations (chaque notation correspond à une entreprise à un instant donné) satisfont l'hypothèse d'échangeabilité requise pour les garanties théoriques de la prédiction conforme. Enfin, le dataset couvre 25 secteurs d'activité différents, offrant une diversité sectorielle représentative du tissu économique américain.

Description détaillée des données

Le dataset comprend 2 031 observations réparties sur 31 variables, sans aucune valeur manquante. Les variables se décomposent en 28 variables numériques (ratios financiers), 2 variables catégorielles (`Sector`, `Rating Agency Name`) et 3 variables d'identification (`Name`, `Symbol`, `Date`) qui seront exclues de l'analyse.

La variable cible `Rating` est une variable catégorielle ordinale représentant la notation de crédit de l'entreprise selon l'échelle traditionnelle. Elle comporte 10 classes allant de AAA (meilleure qualité de crédit) à D (défaut de paiement). La distribution initiale présente un déséquilibre important, avec une forte concentration sur les classes A (671 observations, 33.0%) et BBB (624 observations, 30.7%), tandis que certaines classes sont gravement sous-représentées : D (1 observation), C (2 observations), CC (8 observations) et AAA (52 observations). Ce déséquilibre, avec un ratio maximum/minimum de 671:1, nécessite un regroupement des classes pour garantir la robustesse des modèles.

Les 28 variables explicatives numériques couvrent l'ensemble des dimensions pertinentes pour l'évaluation du risque de crédit d'entreprise, comme synthétisé dans le tableau ci-dessous.

Les deux variables catégorielles complémentaires sont `Sector` (25 secteurs : Technology, Health Care, Consumer Durables, etc.) et `Rating Agency Name` (4 agences : Standard & Poor's, Moody's, Fitch Ratings, Egan-Jones).

Problématique métier et justification

Considérons le scénario d'un analyste financier chargé d'évaluer le risque de crédit d'une entreprise pour décider d'un investissement obligataire. Un modèle de classification classique lui fournirait une prédiction unique : « Cette entreprise est notée BBB ». Cette information, bien qu'utile, reste insuffisante pour une prise de décision éclairée. L'analyste ne dispose d'aucune

[2] `https://www.kaggle.com/datasets/agewerc/corporate-credit-rating`

Catégorie	Variables principales
Liquidité	currentRatio, quickRatio, cashRatio, cashPerShare
Rentabilité	netProfitMargin, returnOnAssets, returnOnEquity, returnOnCapitalEmployed, operatingProfitMargin, grossProfitMargin, pretaxProfitMargin, ebitPerRevenue
Efficacité opérationnelle	assetTurnover, fixedAssetTurnover, daysOfSalesOutstanding, payablesTurnover
Structure financière	debtEquityRatio, debtRatio, companyEquityMultiplier
Flux de trésorerie	freeCashFlowPerShare, operatingCashFlowPerShare, freeCashFlowOperatingCashFlowRatio, operatingCashFlowSalesRatio
Fiscalité & Valorisation	effectiveTaxRate, enterpriseValueMultiple

Table 4: Synthèse des variables du dataset de classification

indication sur la confiance du modèle dans cette prédiction. S'agit-il d'un profil clairement BBB, ou l'entreprise se situe-t-elle à la frontière entre Investment Grade et Speculative Grade ?

La **prédiction conforme pour la classification** enrichit radicalement cette analyse en produisant des **ensembles de prédiction**. Au lieu d'une classe unique, le modèle pourrait indiquer : « Avec 90% de confiance, cette entreprise appartient à l'ensemble {BBB, BB} ». Cette information est beaucoup plus riche pour la décision. Un ensemble contenant uniquement BBB signale une prédiction très confiante et permet une décision rapide. Un ensemble {BBB, BB} indique une entreprise à la frontière Investment Grade/Speculative, justifiant une analyse approfondie. Un ensemble large {A, BBB, BB} révèle une forte incertitude et nécessite une due diligence complète avant investissement.

Dans le contexte réglementaire bancaire, cette approche est particulièrement pertinente. Les accords de Bâle imposent des exigences de fonds propres différentes selon la catégorie de crédit. Une entreprise classée Investment Grade (AAA à BBB) bénéficie de conditions favorables, tandis qu'une notation Speculative Grade (BB et inférieur) entraîne des exigences accrues. La prédiction conforme offre donc une garantie statistique contrôlée : sur 100 prédictions avec un niveau de confiance de 90%, au moins 90 contiendront la vraie classe. Cette propriété est cruciale pour justifier les décisions d'investissement auprès des régulateurs et des comités de risque.

De plus, le regroupement des classes en 6 catégories cohérentes (IG_HIGH, IG_MED, IG_LOW pour Investment Grade ; SPEC_HIGH, SPEC_MED, SPEC_LOW pour Speculative Grade) aligne parfaitement l'approche technique avec les pratiques métier, où la distinction majeure se fait entre obligations investissables et spéculatives.

Analyse exploratoire des données (EDA)

L'analyse exploratoire révèle plusieurs caractéristiques importantes du dataset. La distribution initiale de la variable cible `Rating` présente un déséquilibre critique, avec une concentration massive sur les classes A (33.0%) et BBB (30.7%), représentant ensemble près de 64% des observations. À l'opposé, les classes extrêmes sont gravement sous-représentées : D (1 obs.), C (2 obs.), CC (8 obs.) et AAA (52 obs.), rendant impossible un apprentissage robuste sur ces catégories (Figure 9).

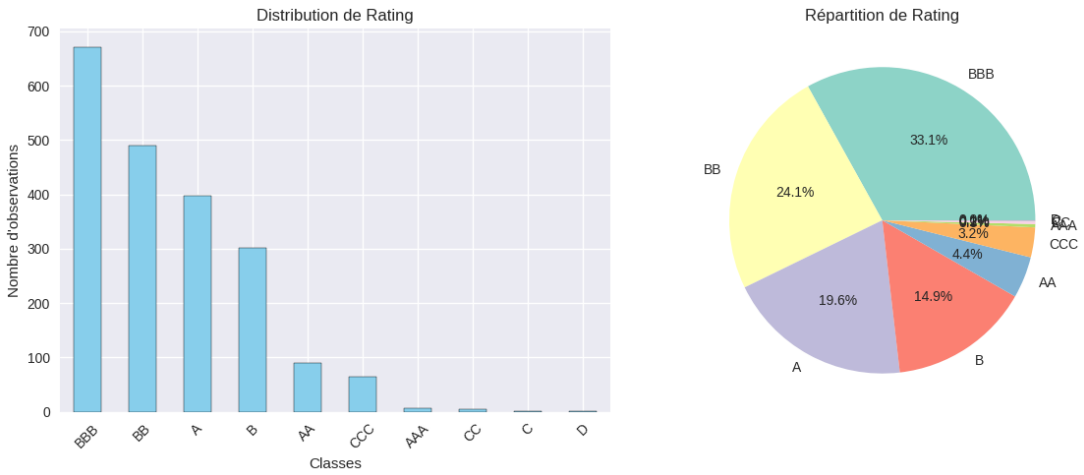

Figure 9: Distribution originale de la variable cible Rating (10 classes fortement déséquilibrées)

Pour remédier à ce déséquilibre, nous avons procédé à un regroupement des classes en 6 catégories basées sur la distinction Investment Grade / Speculative Grade, reflétant les pratiques financières réelles (Figure 10). Cette transformation améliore considérablement l'équilibre du dataset, avec un ratio maximum/minimum passant de 671:1 à seulement 9.3:1, et une classe minimale passant de 1 à 72 observations.

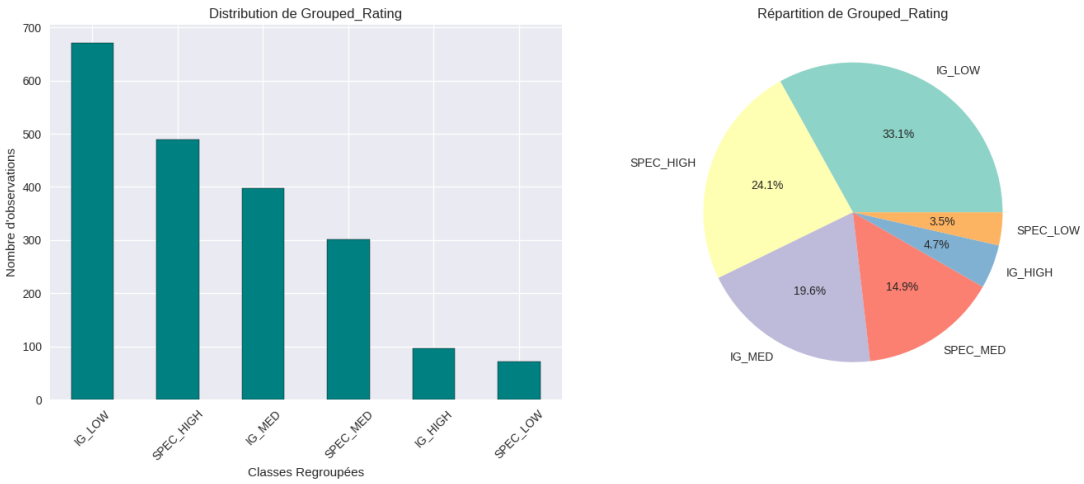

Figure 10: Distribution regroupée de la variable cible (6 classes équilibrées)

L'analyse de corrélation met en évidence plusieurs variables numériques fortement corrélées entre elles. Les indicateurs de rentabilité (`returnOnAssets`, `returnOnEquity`, `returnOnCapitalEmployed`) présentent des corrélations élevées entre eux indiquant que de futurs travaux pourraient envisager d'éliminer certaines redondances (Figure 11).

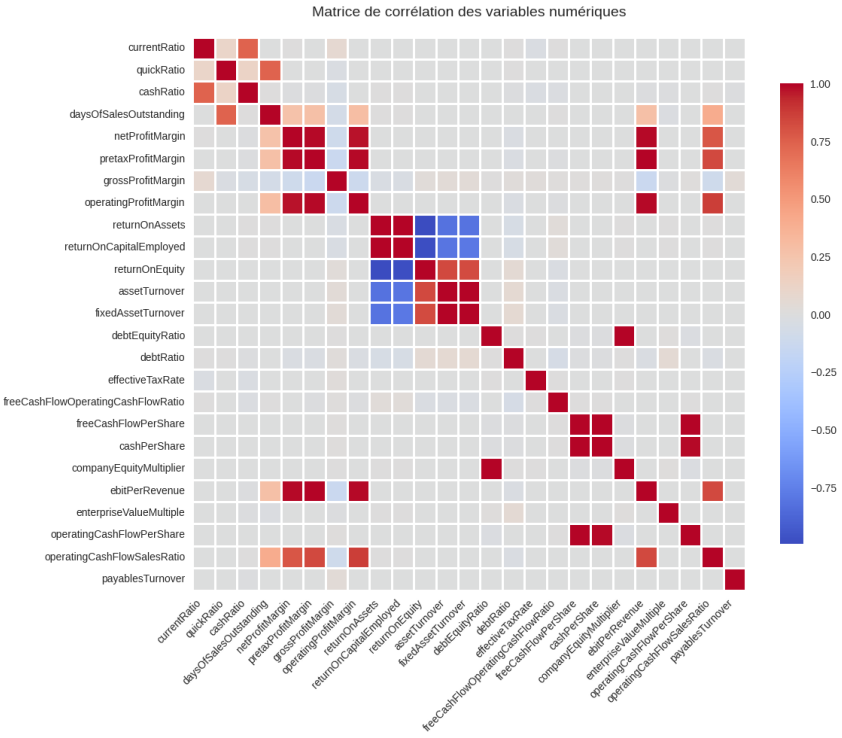

Figure 11: Matrice de corrélation des variables numériques

Enfin, l'analyse de corrélation entre les variables numériques et la classe regroupée encodée (Figure 12) révèle que `debtRatio` (+0.22), `enterpriseValueMultiple` (+0.086) et `cashRatio` (+0.025) sont positivement corrélés avec un meilleur rating, tandis que les indicateurs `payables-Turnover` (-0.06) et `freeCashFlowOperatingCashFlowRatio` (-0.052) montrent une corrélation négative. Toutefois, ces corrélations restent faibles, suggérant que la prédiction de la notation de crédit nécessite une analyse multivariée complexe plutôt qu'une simple relation linéaire avec une ou deux variables.

Figure 12: Corrélations des features avec la variable cible Grouped_Rating

Préparation et prétraitement des données

Le prétraitement des données s'effectue en plusieurs étapes successives. Les variables d'identification (`Name`, `Symbol`, `Date`), ne contenant que des informations nominatives ou temporelles sans valeur prédictive, sont d'abord supprimées du dataset.

La variable cible `Rating` fait l'objet d'un regroupement en 6 classes hiérarchiques pour résoudre le problème de déséquilibre critique :

- **IG_HIGH** (Investment Grade - Haute qualité) : AAA, AA
- **IG_MED** (Investment Grade - Qualité moyenne) : A
- **IG_LOW** (Investment Grade - Qualité satisfaisante) : BBB
- **SPEC_HIGH** (Speculative Grade - Modérément spéculatif) : BB
- **SPEC_MED** (Speculative Grade - Spéculatif) : B
- **SPEC_LOW** (Speculative Grade - Très spéculatif/Défaut) : CCC, CC, C, D

Ce regroupement s'aligne sur les pratiques financières réelles où la frontière majeure se situe entre Investment Grade (BBB et supérieur) et Speculative Grade (BB et inférieur), tout en préservant une granularité suffisante pour distinguer les niveaux de risque au sein de chaque catégorie.

Les variables catégorielles (`Sector`, `Rating Agency Name`) sont ensuite encodées en utilisant un encodage one-hot, créant des variables binaires pour chaque modalité tout en supprimant une modalité de référence pour éviter la multicolinéarité. Les variables numériques sont normalisées par standardisation.

6 Prédiction Conforme pour la Classification

6.1 Choix des méthodes et justification

Nous avons implémenté deux approches de prédiction conforme pour la classification, chacune adaptée à un contexte spécifique.

Split Conformal Prediction (SCP)

Présenté précédemment pour la prédiction conforme en régression, SCP divise les données en trois ensembles disjoints : train, calibration et test. Cette méthode est simple et rapide, adaptée aux datasets volumineux, **ce qui n'est pas le cas ici**. Cela risque de limiter la taille effective du calibration set, impactant la précision des ensembles de prédiction.

Nous aurions voulu utiliser d'autres algorithmes vu en cours comme Jackknife+ ou CV+ pour leur meilleure exploitation d'un faible quantité de données, cependant ils ne semblent pas applicable en classification. C'est pourquoi après quelques recherches, nous nous sommes rabattus sur la Cross-Conformal Prediction (CCP) comme seconde méthode.

Cross-Conformal Prediction (CCP)

CCP exploite la validation croisée K-folds pour maximiser l'utilisation des données disponibles. La méthode a été entièrement implémentée par LLM en s'appuyant sur le papier fondateur de Lei et al. [1]. Elle entraîne K modèles (ici $K = 5$) et combine leurs p-values via moyenne arithmétique pour garantir la validité statistique. Cette implémentation sert de référence pour évaluer les limites de notre SCP face à un dataset de taille modeste.

Note importante : CCP ne rentre pas dans les critères de notation du projet car entièrement générée par IA. Elle constitue uniquement un outil de comparaison pour comprendre l'impact de la taille du calibration set sur les performances de SCP.

Choix des scores de conformité

Deux scores ont été testés pour mesurer la non-conformité d'une prédiction :

Score absolu : $S(X, y) = 1 - \hat{p}(y|X)$ où $\hat{p}(y|X)$ est la probabilité prédite pour la classe y. Simple et interprétable, il pénalise directement les classes avec faible probabilité.

Score cumulatif : $S(X, y) = \sum_{j:\hat{p}(j|X) \geq \hat{p}(y|X)} \hat{p}(j|X)$. Ce score mesure la masse de probabilité cumulée jusqu'à la vraie classe (classement décroissant). Il tend à produire des ensembles plus larges mais mieux calibrés sur les classes difficiles.

6.2 Mise en place des modèles

Division des données et prétraitement

Le preprocessing suit le protocole établi lors de la section précédente. Pour la SCP, split en train (1 219 obs., 60%), calibration (609 obs., 30%) et test (203 obs., 10%). Pour la CCP, fusion du jeu de train et du jeu de calibration (1 828 obs., 90%) avec validation croisée 5-folds, test set identique pour les deux méthodes.

Choix de l'algorithme ML de base

Nous avons utilisé un **RandomForestClassifier** optimisé par GridSearch sur 3-fold CV. Les hyperparamètres testés incluent `n_estimators` $\in \{100, 200, 300\}$, `max_depth` $\in \{5, 8, 10\}$, `min_samples_split`

$\in \{5, 10\}$, et `class_weight` $\in$ {balanced, balanced_subsample}. Malgré cela, la taille du dataset et la complexité du problème ont limité la précision du meilleure modèle à environ 49% d'exactitude sur le test set. Nous avons pu tester une autre grid search sur un `XGBClassifier` mais le gain de performance n'était pas significatif.

Principe de la prédiction conforme en classification

Pour SCP, on calcule les scores de conformité $S_i = S(X_i, y_i)$ sur le calibration set, puis le quantile ajusté :

$$\hat{q} = \text{Quantile}\left(S_1, \ldots, S_n; \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right)$$

L'ensemble de prédiction pour X_{new} inclut toutes les classes y telles que $S(X_{new}, y) \leq \hat{q}$:

$$\hat{C}(X_{new}) = \{y : S(X_{new}, y) \leq \hat{q}\}$$

Pour CCP, la procédure est similaire mais utilise des p-values calculées par fold et combinées par moyenne arithmétique. L'ensemble contient les classes dont la p-value moyenne dépasse α.

6.3 Résultats et interprétation

Synthèse des performances

Sur le test set (203 observations), les quatre configurations testées donnent les résultats suivants :

Méthode	Couverture	Taille moyenne
SCP (score absolu)	87.68%	2.65
SCP (score cumulatif)	85.71%	2.70
CCP (score absolu)	89.16%	2.67
CCP (score cumulatif)	**91.63%**	**2.86**

Table 5: Comparaison des méthodes de prédiction conforme en classification

CCP cumulatif obtient la meilleure couverture (91.63%), proche de la cible théorique de 90%, avec des ensembles de taille moyenne 2.86. Les quatre méthodes présentent des tailles d'ensembles similaires (2.65 à 2.86), indiquant une difficulté comparable à discriminer les classes de rating. Le score cumulatif de CCP surpasse légèrement le score absolu (+2.5 points de couverture), tandis que pour SCP les deux scores donnent des résultats proches en couverture mais le score cumulatif produit des ensembles légèrement plus larges.

Exemples de prédictions

Le tableau ci-dessous illustre les ensembles de prédiction obtenus sur un échantillon aléatoire de 10 observations du test set. Pour chaque observation, on compare les ensembles produits par SCP avec score absolu et score cumulatif.

On observe que les deux scores produisent souvent des ensembles identiques ou très similaires. La vraie classe est systématiquement incluse dans l'ensemble (garantie de couverture respectée localement pour ces exemples). Les ensembles contiennent généralement 2 à 3 classes, reflétant la difficulté de discrimination du modèle. L'observation 78 illustre un cas où le score

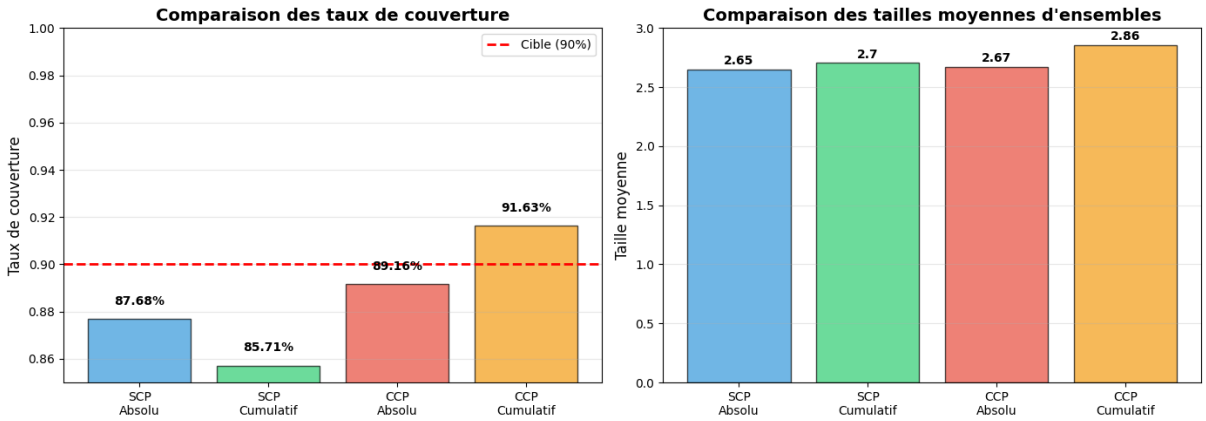

Figure 13: Comparaison des taux de couverture et tailles moyennes d'ensembles pour les quatre configurations

cumulatif produit un ensemble plus large (4 classes) que le score absolu (2 classes), révélant une incertitude accrue pour ce profil particulier.

Limite observée : Les couvertures de SCP (87.68% et 85.71%) restent relativement loin de la cible théorique de 90%, reflétant la difficulté du problème de classification à 6 classes avec un dataset de taille modeste.

Analyse par classe

L'analyse du taux de couverture par classe révèle des disparités importantes. Les classes IG_MED (A) et IG_LOW (BBB), majoritaires dans le dataset (671 et 624 obs. originales), affichent des couvertures supérieures à 90% avec SCP absolu. À l'inverse, les classes minoritaires IG_HIGH et SPEC_LOW présentent des couvertures inférieures à 80%. Cette hétérogénéité reflète le déséquilibre résiduel malgré le regroupement des classes.

En effet, une claire corrélation entre la taille de la classe et la couverture est observée : les classes avec plus d'observations dans le calibration set bénéficient de meilleures estimations des scores de conformité, conduisant à des ensembles plus précis.

Interprétation métier

La prédiction conforme en classification apporte une valeur décisionnelle claire. Un ensemble singleton {IG_MED} signale une entreprise clairement identifiée comme Investment Grade de qualité moyenne, autorisant une décision rapide. Un ensemble {IG_LOW, SPEC_HIGH} révèle une entreprise à la frontière critique BBB/BB, justifiant une analyse approfondie avant décision. Un ensemble large {IG_MED, IG_LOW, SPEC_HIGH} indique une forte incertitude nécessitant une due diligence complète.

Dans le contexte réglementaire de Bâle III, cette information est précieuse afin de pouvoir justifier les décisions d'investissement auprès des régulateurs.

Limites et perspectives

Le principal frein observé est la taille limitée du dataset (2 031 obs.), conduisant à un calibration set de seulement 609 observations pour SCP. Cette contrainte explique les couvertures sous-optimales (87.68% au lieu de 90%). Une solution serait d'enrichir le dataset avec davantage

Obs	Vraie classe	SCP Absolu	SCP Cumulatif
15	IG_LOW	IG_LOW, SPEC_HIGH, SPEC_MED	IG_LOW, SPEC_HIGH, SPEC_MED
9	IG_LOW	IG_HIGH, IG_LOW, IG_MED	IG_HIGH, IG_LOW, IG_MED
115	IG_MED	IG_HIGH, IG_MED	IG_HIGH, IG_MED
78	SPEC_MED	SPEC_HIGH, SPEC_MED	IG_LOW, SPEC_HIGH, SPEC_LOW, SPEC_MED
66	IG_MED	IG_LOW, IG_MED	IG_LOW, IG_MED, SPEC_HIGH
45	SPEC_HIGH	IG_LOW, IG_MED, SPEC_HIGH	IG_LOW, IG_MED, SPEC_HIGH
143	IG_MED	IG_LOW, IG_MED	IG_LOW, IG_MED, SPEC_HIGH
177	SPEC_HIGH	IG_LOW, IG_MED, SPEC_HIGH	IG_LOW, IG_MED, SPEC_HIGH
200	IG_LOW	IG_LOW, IG_MED, SPEC_HIGH	IG_LOW, IG_MED, SPEC_HIGH
180	IG_LOW	IG_LOW, IG_MED, SPEC_HIGH	IG_LOW, IG_MED, SPEC_HIGH

Table 6: Exemples d'ensembles de prédiction pour SCP avec les deux scores

d'entreprises notées, ou d'appliquer des techniques d'augmentation de données pour renforcer les classes minoritaires.

7 Conclusion

Ce projet a implémenté et comparé la **régression quantile** et la **prédiction conforme** sur deux problématiques financières : l'évaluation du risque de crédit individuel (régression) et la notation d'entreprises (classification).

Dans le secteur financier, ces techniques apportent trois avantages majeurs : (1) la **quantification de l'incertitude** permet d'adapter les conditions de prêt selon le profil, (2) les **garanties statistiques formelles** répondent aux exigences réglementaires de Bâle III, (3) la **détection automatique des cas ambigus** optimise l'allocation des ressources d'analyse entre traitement automatisé et expertise humaine.

Les perspectives incluent l'implémentation de Conformal Quantile Regression pour des intervalles adaptatifs, l'enrichissement des datasets et l'exploration de modèles de base plus performants.

References

[1] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.