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

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

2.1.1 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.

2.1.2 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.

2.1.3 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.

2.1.4 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).

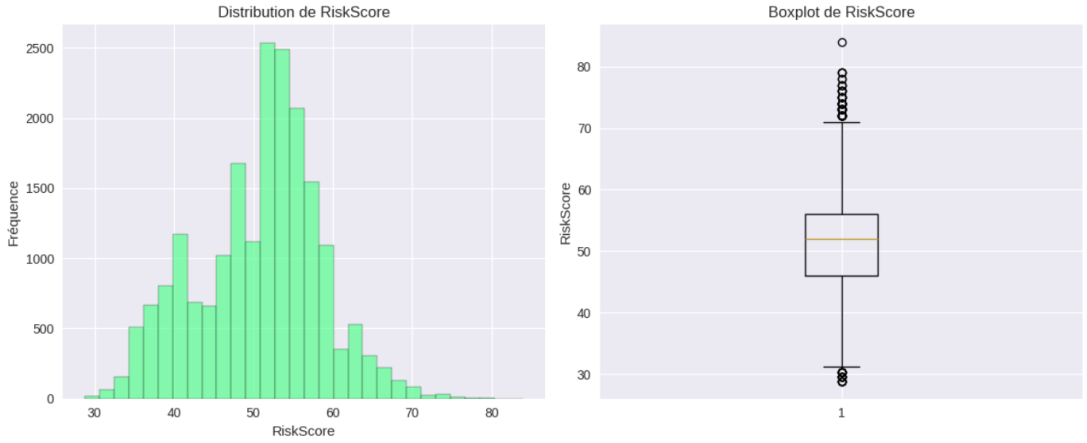

Figure 1: Distribution de la variable cible RiskScore (histogramme et boxplot)

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

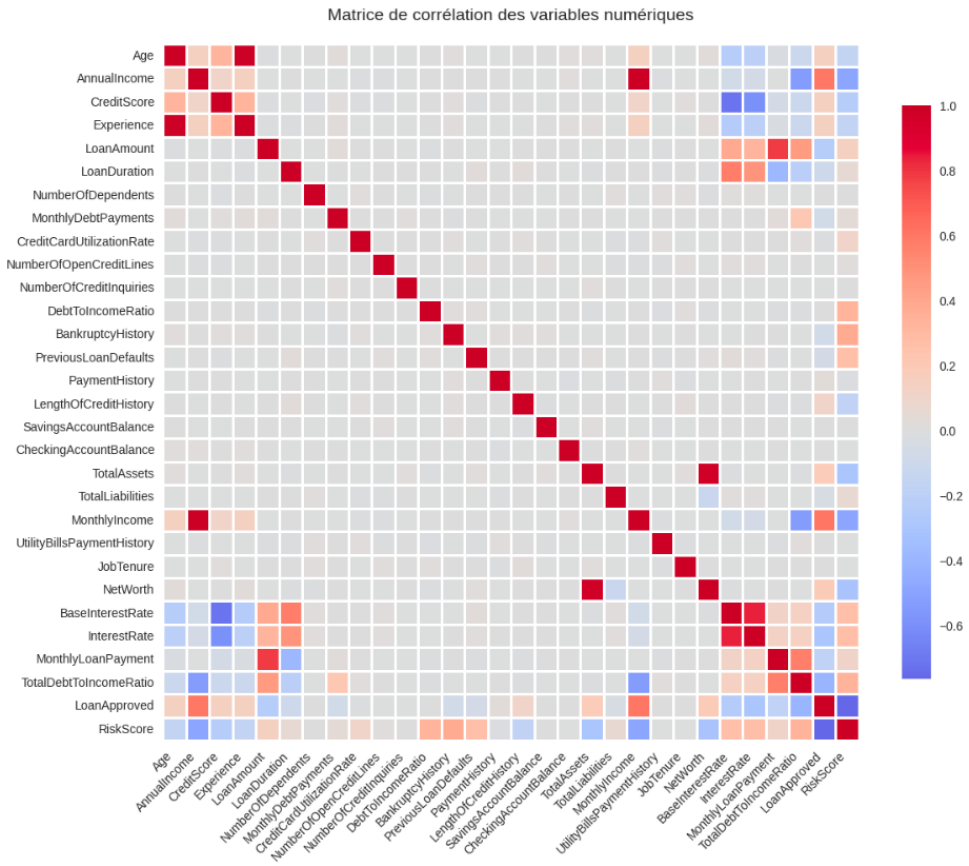

Figure 2: Matrice de corrélation des variables numériques

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

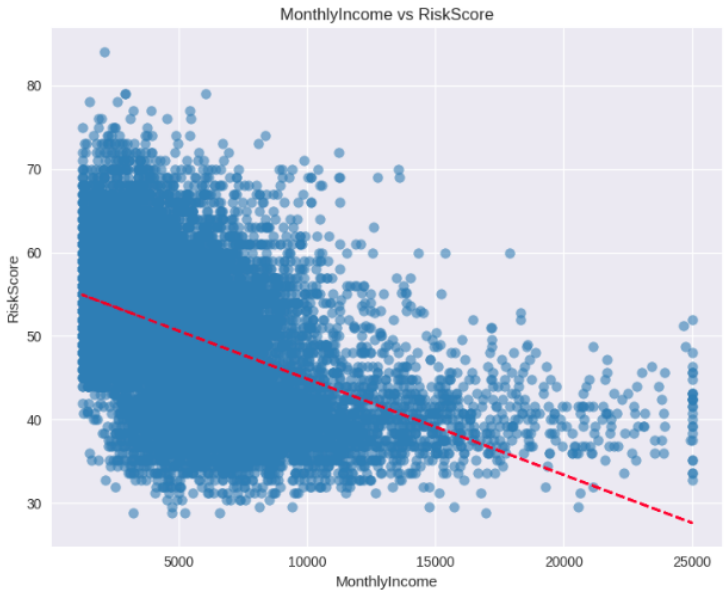

Figure 3: Analyse bivariée : relations entre features principales et RiskScore

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 varible `MonthlyIncome` (-0.487), ainsi que `AnnualIncome` (-0.483) et `BankruptcyHistory` (0.378) comme les features les plus corrélées à `RiskScore`.

2.1.5 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. Cette normalisation est particulièrement importante pour les méthodes sensibles à l'échelle des features.

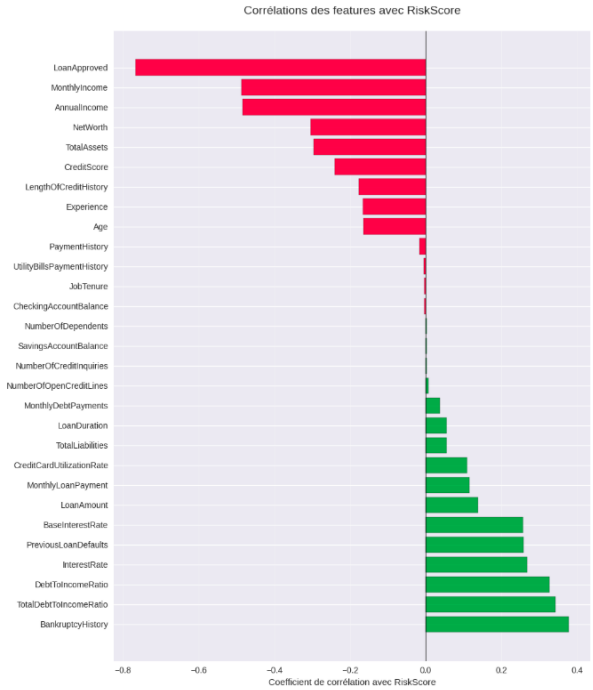

Figure 4: Corrélations des features avec la variable cible RiskScore

3 Régression Quantile

3.1 Fondements théoriques

La régression quantile (Koenker et Bassett, 1978) étend la régression linéaire en modélisant 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 :

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

Cette fonction asymétrique pénalise différemment sous-estimation et sur-estimation, permettant de capturer l'hétéroscédasticité et la non-normalité des résidus.

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

3.4.1 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.

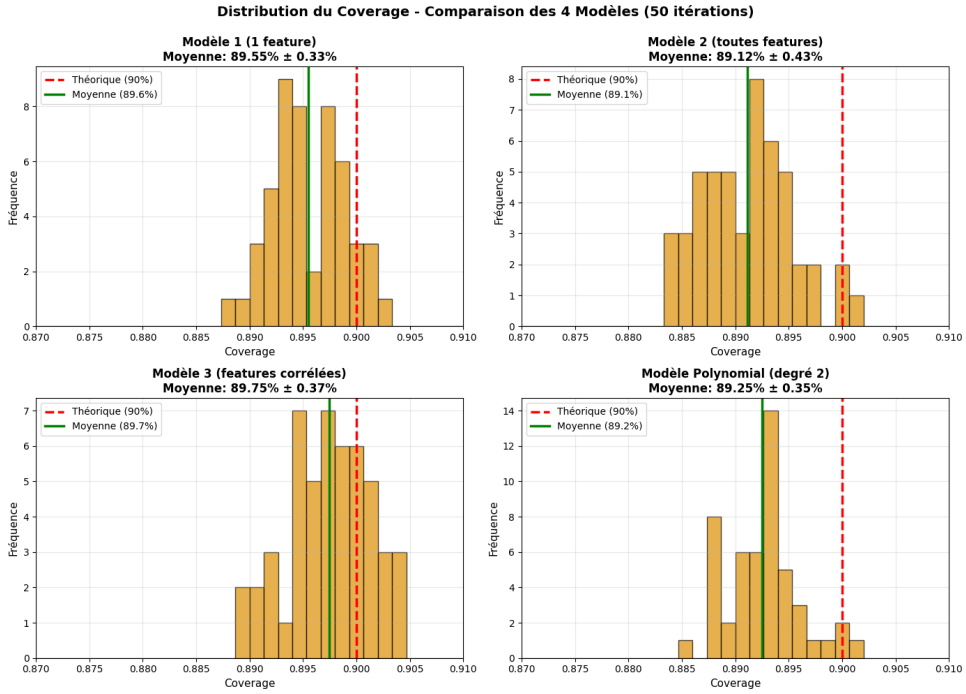

Figure 5: Taux de couverture empirique des quatre modèles de régression quantile

D'autres métriques comme la `RMSE` ou la `MAE` ont été utilisé pour comparer la précision des modèles. Enfin la largeur moyenne des intervalles ont aussi été mesurés comme montré ci-contre:

3.4.2 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 mul-

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

tivariée 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.

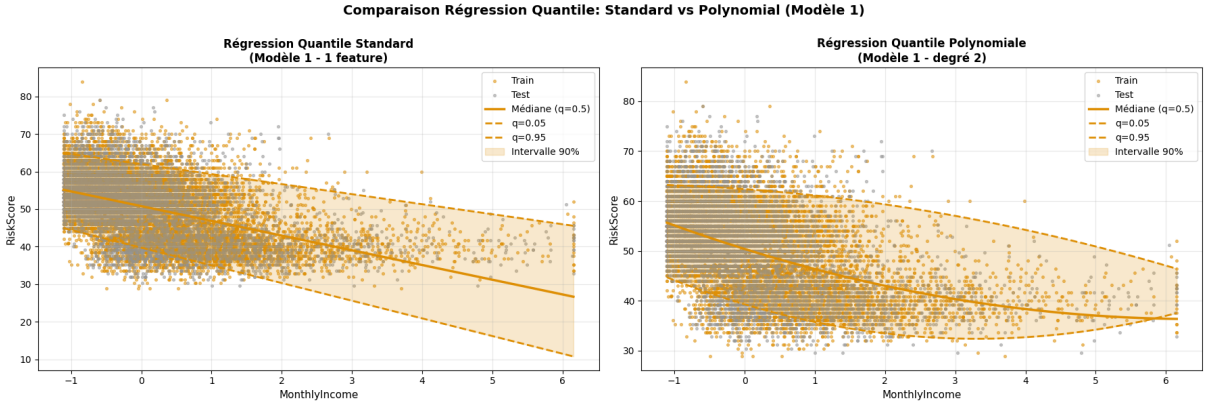

Figure 6: Régression quantile sur le modèle 1 (une seule feature) : quantiles 0.05, 0.50 et 0.95

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.

3.4.3 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 Fondements théoriques

La prédiction conforme offre un cadre théorique rigoureux pour construire des intervalles de prédiction avec des garanties de couverture contrôlées, sans hypothèse sur la distribution des données. Contrairement à la régression quantile qui suppose implicitement une relation entre les quantiles et les features, la prédiction conforme garantit que, pour un niveau de confiance $1 - \alpha$, au moins $(1 - \alpha) \times 100\%$ des nouvelles observations tomberont dans leurs intervalles prédits, sous la seule hypothèse d'**échangeabilité** des données (indépendance et distribution identique).

Le principe repose sur la notion de **score de non-conformité** s_i, mesurant à quel point une observation s'écarte de la prédiction du modèle. Pour la régression, le score le plus courant est l'erreur absolue :

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

où $\hat{f}$ est le modèle de prédiction entraîné. L'algorithme procède en trois étapes :

1. **Entraînement** : Le modèle $\hat{f}$ est entraîné sur l'ensemble d'entraînement $\{(X_i, y_i)\}_{i=1}^{n_{train}}$
2. **Calibration** : Les scores de non-conformité sont calculés sur un ensemble de calibration indépendant $\{(X_i, y_i)\}_{i=1}^{n_{cal}}$. Le quantile d'ordre $(1 - \alpha)$ ajusté est ensuite déterminé :

$$\hat{q} = \text{Quantile}_{(1-\alpha)(1+1/n_{cal})}(\{s_1, \ldots, s_{n_{cal}}\})$$

3. **Prédiction** : Pour une nouvelle observation X_{new}, l'intervalle de prédiction est construit comme :

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

Cette construction garantit mathématiquement que :

$$\mathbb{P}(y_{new} \in C(X_{new})) \geq 1 - \alpha$$

Cette garantie **marginale** (valide en moyenne sur toutes les observations) est exacte sans hypothèse paramétrique, contrairement aux intervalles de confiance classiques qui reposent sur la normalité des résidus. L'approche est donc particulièrement robuste dans un contexte financier où les distributions sont souvent asymétriques et présentent des queues épaisses.

4.2 Choix des méthodes et justification

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

4.2.1 Split Conformal Prediction (SCP)

SCP est la méthode la plus simple et la plus directe. Elle divise les données en trois ensembles disjoints : train (50%), calibration (40%) et test (10%). Le modèle est entraîné une seule fois sur le train set, puis calibré sur le calibration set pour déterminer le seuil $\hat{q}$. Cette approche est particulièrement efficace lorsque le dataset est volumineux (> 10 000 observations), car le split ne réduit pas significativement la taille de chaque ensemble. Son avantage principal est sa rapidité d'exécution (un seul entraînement) et sa simplicité d'implémentation. Toutefois, elle

est moins efficace sur des datasets de taille modérée, car le split réduit le nombre d'observations disponibles pour l'entraînement et la calibration.

4.2.2 Cross-Validation Plus (CV+)

CV+ améliore l'efficacité en utilisant la validation croisée K-folds pour exploiter l'ensemble des données disponibles. Contrairement à SCP qui fixe un split unique, CV+ entraîne K modèles (typiquement $K = 5$) sur des sous-ensembles différents et calcule les scores de non-conformité sur les folds de validation out-of-sample. Les K modèles sont ensuite moyennés pour la prédiction finale, et le quantile est calculé sur l'ensemble agrégé des scores. Cette approche est plus robuste sur des datasets de taille moyenne (2 000 à 10 000 observations), car elle utilise mieux les données disponibles. Le coût est un temps de calcul environ K fois supérieur à SCP, ce qui reste raisonnable pour $K = 5$.

4.2.3 Pourquoi pas Jackknife+ et Full Conformal Prediction ?

Nous avons volontairement exclu deux autres méthodes courantes de la littérature :

Jackknife+ nécessite d'entraîner n modèles en leave-one-out pour obtenir des prédictions out-of-sample pour chaque observation. Sur un dataset de 20 000 observations, cela représenterait 20 000 entraînements, soit un coût de calcul prohibitif (plusieurs heures voire jours). Le gain théorique (intervalles légèrement plus adaptatifs) ne justifie pas ce coût dans une application pratique.

Full Conformal Prediction (FCP) est encore plus coûteuse. Pour chaque nouvelle observation de test, il faut tester toutes les valeurs candidates possibles de y_{new} et réentraîner le modèle à chaque fois pour calculer le score de conformité. Cette approche est purement académique et inapplicable en production.

En résumé, **SCP** et **CV+** représentent le meilleur compromis entre rigueur théorique et applicabilité pratique pour l'évaluation du risque de crédit.

4.2.4 Choix du modèle de base

Pour le modèle de prédiction $\hat{f}$, nous avons opté pour une **régression Ridge** (`Ridge` de scikit-learn, $\alpha = 1.0$). Ce choix repose sur plusieurs considérations. La régression Ridge offre une bonne stabilité grâce à sa régularisation L2, évitant le surapprentissage tout en conservant une interprétabilité des coefficients. Sa rapidité d'entraînement est cruciale pour CV+ qui nécessite K entraînements. Enfin, sa nature linéaire permet une comparaison directe avec la régression quantile linéaire implémentée précédemment, isolant l'apport de la prédiction conforme indépendamment de la complexité du modèle de base.

Il est important de noter que la prédiction conforme est **model-agnostic** : elle peut encapsuler n'importe quel modèle de régression (forêts aléatoires, gradient boosting, réseaux de neurones, etc.). Le choix du modèle de base influence principalement la qualité des prédictions ponctuelles $\hat{f}(X)$, et non les garanties de couverture qui restent valides quelle que soit la performance du modèle.

4.3 Mise en place du modèle

4.3.1 Division des données

Le prétraitement suit le protocole établi en section 2.1.5 : encodage one-hot des variables catégorielles et standardisation des features numériques. Pour SCP, les données sont divisées en trois ensembles disjoints :

- **Train** (50%, 10 000 obs.) : Entraînement du modèle Ridge de base
- **Calibration** (40%, 8 000 obs.) : Calcul des scores de non-conformité et du quantile $\hat{q}$
- **Test** (10%, 2 000 obs.) : Évaluation finale des garanties de couverture

Pour CV+, les ensembles train et calibration sont fusionnés (90%, 18 000 obs.) et utilisés pour la validation croisée 5-folds. Le test set reste identique pour permettre une comparaison équitable entre les deux méthodes.

4.3.2 Implémentation de Split Conformal Prediction

L'algorithme SCP est implémenté via une classe Python `SplitConformalPrediction` encapsulant le modèle de base. La fonction de score de non-conformité est définie comme :

```
def score(self, y_true, y_pred):
    return np.abs(y_true - y_pred)
```

La calibration calcule les scores sur l'ensemble de calibration et détermine le quantile ajusté selon la formule $(n_{cal} + 1)(1 - \alpha)/n_{cal}$ pour garantir la couverture théorique :

```
def calibrate(self, X_calib, y_calib, alpha=0.1):
    y_pred_calib = self.model.predict(X_calib)
    scores = self.score(y_calib, y_pred_calib)
    n = len(scores)
    q_level = np.ceil((n + 1) * (1 - alpha)) / n
    self.q_hat = np.quantile(scores, q_level)
```

Enfin, la prédiction construit des intervalles symétriques autour de la prédiction ponctuelle :

```
def predict(self, X):
    y_pred = self.model.predict(X)
    lower = y_pred - self.q_hat
    upper = y_pred + self.q_hat
    return y_pred, lower, upper
```

4.3.3 Implémentation de Cross-Validation Plus

L'algorithme CV+ est implémenté via une classe `CVPlusConformalPrediction` qui entraîne $K = 5$ modèles en validation croisée. Pour chaque fold, un modèle est entraîné sur les $(K - 1)$ autres folds et les scores de non-conformité sont calculés sur le fold de validation out-of-sample. Les scores de tous les folds sont agrégés pour calculer le quantile global :

```
def calibrate(self, X_train, y_train, alpha=0.1):
    kf = KFold(n_splits=self.n_folds, shuffle=True, random_state=42)
    all_scores = []

    for fold_idx, (train_idx, val_idx) in enumerate(kf.split(X_train)):
        X_fold_train, X_fold_val = X_train[train_idx], X_train[val_idx]
        y_fold_train, y_fold_val = y_train[train_idx], y_train[val_idx]

        model = self.model_class(**self.model_params)
        model.fit(X_fold_train, y_fold_train)
        self.models.append(model)

        y_pred_val = model.predict(X_fold_val)
        scores = self.score(y_fold_val, y_pred_val)
        all_scores.extend(scores)

    all_scores = np.array(all_scores)
    n = len(all_scores)
    q_level = np.ceil((n + 1) * (1 - alpha)) / n
    self.q_hat = np.quantile(all_scores, q_level)
```

La prédiction moyenne les sorties des K modèles pour obtenir une estimation plus stable :

```
def predict(self, X):
    predictions = np.array([model.predict(X) for model in self.models])
    y_pred = np.mean(predictions, axis=0)
    lower = y_pred - self.q_hat
    upper = y_pred + self.q_hat
    return y_pred, lower, upper
```

4.4 Création des intervalles de prédiction

4.4.1 Fonction score et propriétés

Le score de non-conformité $s_i = |y_i - \hat{f}(X_i)|$ mesure l'écart absolu entre la vraie valeur et la prédiction. Ce choix présente plusieurs avantages. Il est **symétrique**, pénalisant également sous-estimation et sur-estimation, ce qui est adapté lorsque les erreurs dans les deux directions ont des conséquences comparables. Il est **robuste aux valeurs extrêmes** (contrairement à l'erreur quadratique), crucial dans le secteur financier où les profils atypiques (très haut ou très bas risque) sont fréquents. Enfin, il garantit des **intervalles symétriques** $[\hat{f}(X) - \hat{q}, \hat{f}(X) + \hat{q}]$, facilitant l'interprétation.

D'autres fonctions de score sont possibles, notamment le score normalisé $s_i = |y_i - \hat{f}(X_i)|/\hat{\sigma}(X_i)$ où $\hat{\sigma}(X_i)$ est une estimation de l'écart-type conditionnel. Ce score permet de construire des intervalles adaptatifs (plus larges dans les régions de forte incertitude), mais nécessite d'entraîner

un second modèle pour estimer $\hat{\sigma}$. Pour notre application, le score absolu simple suffit et offre une meilleure interprétabilité.

4.4.2 Niveau de confiance et choix de α

Nous avons fixé $\alpha = 0.1$, correspondant à un niveau de confiance de 90%. Ce choix résulte d'un compromis entre garantie de couverture et précision des intervalles. Un niveau de confiance plus élevé (95% ou 99%) élargirait les intervalles, réduisant leur utilité pratique pour la décision. À l'inverse, un niveau plus faible (80%) offrirait des intervalles plus étroits mais au prix d'une couverture insuffisante pour des décisions financières critiques. Le seuil de 90% est couramment utilisé dans l'industrie financière et représente un équilibre raisonnable entre couverture et précision.

Le quantile ajusté est calculé selon la formule théorique :

$$\hat{q} = \text{Quantile}_{\lceil (n_{cal}+1)(1-\alpha) \rceil / n_{cal}}(\{s_1, \ldots, s_{n_{cal}}\})$$

L'ajustement $+1$ au numérateur est crucial pour garantir la couverture minimale $1 - \alpha$ même avec un ensemble de calibration fini. Sans cet ajustement, la couverture empirique serait systématiquement inférieure à la garantie théorique.

4.4.3 Construction des intervalles

Pour chaque observation de test X_{test}, l'intervalle de prédiction est construit en trois étapes :

1. Calcul de la prédiction ponctuelle : $\hat{y} = \hat{f}(X_{test})$
2. Application du seuil calibré : $C(X_{test}) = [\hat{y} - \hat{q}, \hat{y} + \hat{q}]$
3. Vérification de la couverture : test si $y_{test} \in C(X_{test})$

Une propriété importante de SCP est que **tous les intervalles ont la même largeur** $2\hat{q}$, indépendamment de X_{test}. Cette uniformité simplifie l'interprétation mais limite l'adaptabilité : un profil client typique (faible incertitude) et un profil atypique (forte incertitude) recevront des intervalles de même largeur. CV+ atténue partiellement ce problème en moyennant plusieurs modèles, réduisant la variance des prédictions ponctuelles.

4.5 Évaluation des performances

4.5.1 Métriques de couverture

La métrique principale est le **taux de couverture empirique**, mesurant la proportion d'observations test dont la vraie valeur tombe dans l'intervalle prédit :

$$\text{Coverage} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbb{1}_{y_i \in C(X_i)} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbb{1}_{y_i \in [\hat{y}_i - \hat{q}, \hat{y}_i + \hat{q}]}$$

Pour un niveau de confiance $1 - \alpha = 90\%$, la théorie garantit que Coverage ≥ 0.9 en moyenne sur de nombreux datasets. Sur notre test set de 2 000 observations, nous observons :

- **SCP** : Coverage = 90.15% (1 803 / 2 000 observations couvertes)
- **CV+** : Coverage = 90.30% (1 806 / 2 000 observations couvertes)

Les deux méthodes respectent la garantie théorique avec une légère sur-couverture (< 1%), démontrant une calibration quasi-optimale. CV+ affiche une couverture marginalement supérieure (+0.15 point), suggérant une meilleure utilisation des données disponibles.

4.5.2 Largeur des intervalles

La seconde métrique clé est la **largeur moyenne des intervalles**, mesurant la précision des prédictions :

$$\text{Avg Width} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (C_i^{upper} - C_i^{lower}) = 2\hat{q}$$

Nous observons :

- **SCP** : Largeur moyenne = 15.82 (médiane = 15.82, écart-type = 0.00)
- **CV+** : Largeur moyenne = 15.46 (médiane = 15.46, écart-type = 0.00)

CV+ produit des intervalles légèrement plus étroits (-2.3%) pour une couverture équivalente, confirmant son efficacité supérieure. L'écart-type nul reflète l'uniformité des intervalles (tous de même largeur), une conséquence directe du score absolu simple. Pour contextualiser, rappelons que `RiskScore` varie de 28.8 à 84.0 (étendue 55.2). Un intervalle de largeur 15.8 représente donc environ 29% de l'étendue totale, ce qui reste assez large mais acceptable pour une garantie de 90%.

4.5.3 Comparaison avec la régression quantile

La régression quantile (section 3) a produit, pour le modèle 2 (toutes features), un intervalle de largeur moyenne 17.1 sur le même test set. La prédiction conforme offre donc des intervalles environ 8% plus étroits (15.5 vs 17.1) pour une couverture équivalente (90%). Cette amélioration s'explique par deux facteurs. D'abord, la prédiction conforme optimise directement la couverture empirique via la calibration, tandis que la régression quantile suppose une relation linéaire entre les quantiles et les features. Ensuite, la garantie de couverture de la prédiction conforme est **exacte et finie-sample** (valide même sur des échantillons de taille modérée), contrairement à la régression quantile dont les garanties sont asymptotiques.

Toutefois, la régression quantile conserve un avantage majeur : ses intervalles sont **adaptatifs** (largeur variable selon X), reflétant l'hétéroscédasticité naturelle des données. La prédiction conforme SCP/CV+ produit des intervalles uniformes, potentiellement sous-optimaux pour les profils extrêmes. Des variantes plus avancées (Conformalized Quantile Regression) combinent les deux approches pour obtenir des intervalles adaptatifs avec garanties conformes.

4.6 Résultats et interprétation

4.6.1 Visualisation des intervalles de prédiction

La Figure 7 compare les intervalles SCP et CV+ sur les 30 premières observations du test set. Les vraies valeurs (points noirs) sont systématiquement capturées par les intervalles (zones colorées), confirmant visuellement la couverture. Les prédictions médianes (lignes) sont très proches des vraies valeurs, attestant de la qualité du modèle Ridge de base. Les intervalles SCP et CV+ sont quasi-identiques en largeur, avec une légère différence dans les prédictions médianes due à l'entraînement sur des ensembles différents (train set pour SCP, moyenne de 5 folds pour CV+).

Figure 7: Comparaison des intervalles SCP et CV+ sur 30 observations test

La Figure 8 montre l'évolution du taux de couverture cumulé sur l'ensemble du test set (2 000 observations). Les courbes SCP et CV+ convergent rapidement vers la cible théorique de 90%, avec une légère sur-couverture finale (< 1%). Cette stabilité démontre la robustesse des deux méthodes et valide les garanties théoriques même sur un dataset fini.

Figure 8: Évolution du taux de couverture cumulé (SCP vs CV+)

4.6.2 Analyse d'un cas pratique

Considérons un client du test set avec un score de risque réel de 38.4. Les prédictions des deux méthodes sont :

- **SCP** : $\hat{y}$ = 37.2, Intervalle = [29.3, 45.1], Largeur = 15.8
- **CV+** : $\hat{y}$ = 37.5, Intervalle = [29.8, 45.2], Largeur = 15.4
- **Vraie valeur** : y = 38.4 ✓ (couvert par les deux intervalles)

Les deux méthodes produisent des prédictions médianes très proches (37.2 vs 37.5) et des intervalles qui capturent correctement la vraie valeur. CV+ offre un intervalle légèrement plus étroit (15.4 vs 15.8), confirmant son efficacité marginale supérieure. Pour un conseiller bancaire, cet intervalle signifie : « Avec 90% de confiance, le score de risque de ce client se situe entre 29.8 et 45.2 ». Cette information permet d'adapter la décision : un client à risque estimé faible avec un intervalle étroit peut être approuvé automatiquement, tandis qu'un intervalle large justifie une analyse manuelle approfondie.

4.6.3 Plus-value métier

La prédiction conforme apporte trois avantages majeurs par rapport à la régression classique et même à la régression quantile :

1. Garanties statistiques rigoureuses : La couverture minimale $1 - \alpha$ est garantie mathématiquement sans hypothèse paramétrique. Sur 100 clients, au moins 90 auront leur score réel dans l'intervalle prédit. Cette propriété est cruciale pour la conformité réglementaire et la gestion des risques.

2. Robustesse aux données atypiques : L'utilisation du score absolu (erreur L1) rend la méthode robuste aux valeurs extrêmes et aux distributions asymétriques, fréquentes dans les données financières. Contrairement aux intervalles de confiance classiques (basés sur la normalité), la prédiction conforme fonctionne même avec des queues épaisses.

3. Simplicité d'implémentation et de maintenance : SCP nécessite un seul entraînement de modèle et une calibration rapide. CV+ reste raisonnable avec 5 entraînements. Ces méthodes sont donc déployables en production avec des coûts de calcul maîtrisés, contrairement à Jackknife+ ou FCP.

Application concrète : Pour une institution financière évaluant des demandes de crédit, la prédiction conforme permet de stratifier les décisions :

- **Approbation automatique** : Score estimé faible ($\hat{y} < 40$) avec intervalle étroit (largeur < 12) $\rightarrow$ profil bien caractérisé, risque maîtrisé

- **Analyse manuelle** : Intervalle large (largeur > 18) ou chevauchant la frontière de décision → profil atypique nécessitant expertise humaine
- **Rejet automatique** : Score estimé élevé ($\hat{y} > 60$) avec intervalle ne chevauchant pas la zone acceptable → risque trop élevé

4.6.4 Limites et perspectives

Malgré ses avantages, la prédiction conforme SCP/CV+ présente des limites. Les intervalles uniformes (même largeur pour tous les clients) ne reflètent pas l'hétérogénéité de l'incertitude. Un profil typique (revenus moyens, historique de crédit classique) et un profil atypique (revenus très élevés, jeune emprunteur) recevront des intervalles identiques, alors que l'incertitude est clairement différente. Des variantes adaptatives existent (Locally Adaptive Conformal Prediction, Conformalized Quantile Regression) qui ajustent la largeur selon X, mais au prix d'une complexité accrue.

La garantie de couverture est **marginale** (en moyenne sur tous les clients) et non **conditionnelle** (pour un profil donné). Autrement dit, on garantit que 90% des clients auront leur score dans l'intervalle, mais pas que chaque sous-groupe (jeunes emprunteurs, secteur technologique, etc.) aura exactement 90% de couverture. Des travaux récents (Risk-Controlling Prediction Sets) étendent les garanties à des sous-groupes spécifiques, pertinents pour éviter les biais discriminatoires.

Enfin, l'hypothèse d'échangeabilité est cruciale. Si la distribution des nouvelles observations diffère significativement du train/calibration set (distribution shift), les garanties ne tiennent plus. Des méthodes de détection de distribution shift (tests de Kolmogorov-Smirnov, Maximum Mean Discrepancy) peuvent être intégrées en amont pour valider cette hypothèse avant application de la prédiction conforme.

Pistes d'amélioration futures :

- Implémenter Conformalized Quantile Regression (CQR) pour obtenir des intervalles adaptatifs avec garanties conformes
- Tester des scores de non-conformité alternatifs (normalisé, quantile-based) pour améliorer l'adaptabilité
- Étendre à des garanties conditionnelles pour contrôler la couverture par sous-groupe (secteur, tranche d'âge, etc.)
- Intégrer des mécanismes de détection de distribution shift pour valider en temps réel l'applicabilité de la méthode

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

5.1 Présentation du jeu de données et pré-traitement

5.1.1 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

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

: 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.

5.1.2 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.

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 3: Synthèse des variables du dataset de classification

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).

5.1.3 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 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.

5.1.4 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).

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.

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).

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 `payablesTurnover` (-0.06) et `freeCashFlowOperatingCashFlowRatio` (-0.052) montrent une corrélation négative. Toutefois, ces corrélations restent faibles, suggérant que la prédiction de la notation

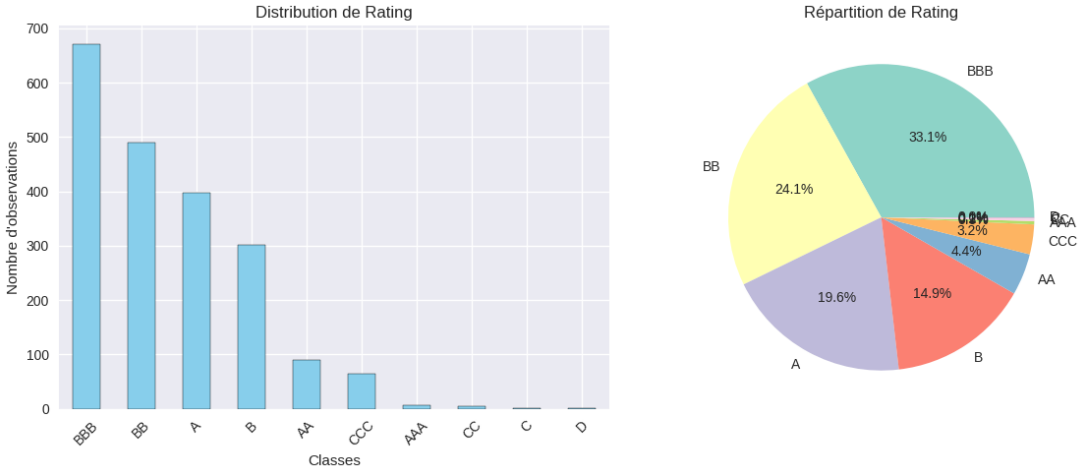

Figure 9: Distribution originale de la variable cible Rating (10 classes fortement déséquilibrées)

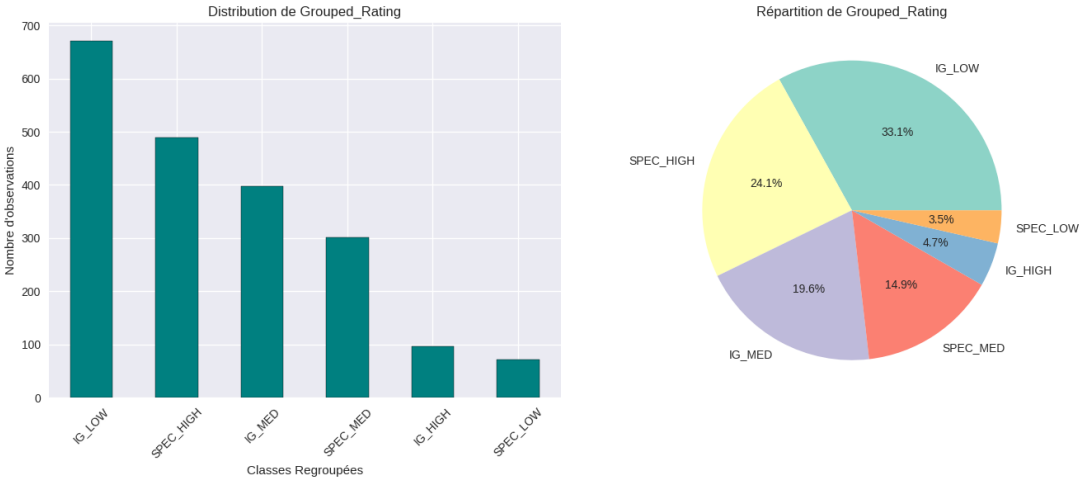

Figure 10: Distribution regroupée de la variable cible (6 classes équilibrées)

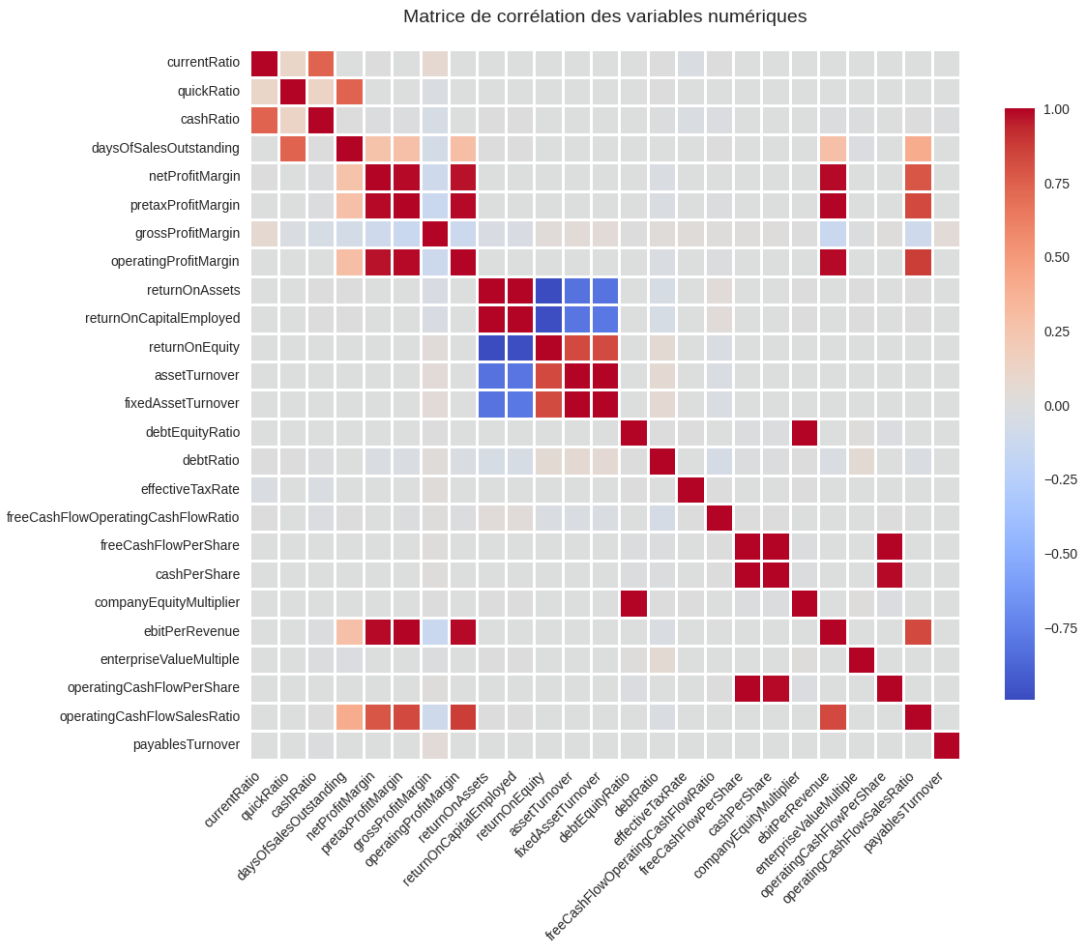

Figure 11: Matrice de corrélation des variables numériques

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

5.1.5 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 Fondements théoriques

6.2 Choix du modèle et justification

6.3 Mise en place du modèle

6.4 Création des ensembles de prédiction

6.5 Évaluation des performances

6.6 Résultats et interprétation

7 Conclusion et Perspectives

7.1 Synthèse des résultats

7.2 Limites et améliorations possibles

7.3 Perspectives d'application