Exploration de la Prédiction Conforme dans les Modèles de Régression et de Classification

Études sur le marché publicitaire et la Reconnaissance des images.

Lama Zidan
MASTER 2 MIASHS
UFR Anthropologie Sociologie Science Politique (ASSP)
Université de Lyon, Université Lumière Lyon 2
Enseignant: Rémi VAUCHER

17 December 2023

plan

1 Introduction

1.1 Présentation Générale du Projet

Dans un monde caractérisé par une complexité croissante des données, l'importance d'approches statistiques sophistiquées et robustes devient cruciale. Ce projet explore des méthodologies avancées en statistique et en apprentissage automatique, avec un accent particulier sur trois domaines clés : la régression quantile, la prédiction conforme appliquée à la régression, et la prédiction conforme dans le cadre de la classification. Ces méthodes sont des outils essentiels pour comprendre et prédire les comportements complexes des données dans divers contextes, offrant des perspectives enrichissantes dans le domaine de l'analyse statistique.

1.2 Objectifs du Projet

L'objectif de ce projet est double : premièrement, fournir une analyse critique des trois méthodes statistiques mentionnées, en soulignant leurs forces et leurs limites ; deuxièmement, illustrer leurs applications pratiques. Nous cherchons à démontrer comment ces approches peuvent améliorer la précision et la fiabilité des modèles prédictifs, en particulier dans des situations où les données sont hétérogènes et présentent des anomalies. L'accent est mis sur l'exploration de leur potentiel dans des scénarios réels, avec un intérêt particulier pour les situations où l'interprétation et la quantification de l'incertitude sont cruciales.

1.3 Brève Description des Trois Domaines Principaux

• Régression Quantile : Cette méthode transcende la régression moyenne traditionnelle en modélisant différents quantiles de la distribution des données. Elle est particulièrement pertinente pour l'analyse des comportements aux extrémités de la distribution, offrant ainsi une vision plus nuancée et détaillée des tendances des données.

• Prédiction Conforme sur la Régression : Basée sur des principes statistiques rigoureux, cette approche moderne quantifie l'incertitude dans les prédictions de régression. Nous explorerons des techniques telles que le "jackknife+" et le "cross-validation+", fournissant un cadre robuste et fiable pour l'évaluation des prédictions.

• Prédiction Conforme sur la Classification : L'application de la prédiction conforme, spécifiquement le "split conformal prediction", à la classification permet une meilleure compréhension et quantification de la confiance dans les prédictions de classification. Cette approche est d'une importance capitale dans les domaines nécessitant des décisions prudentes et bien informées.

2 Régression Quantile

Dans cette section de notre étude, nous adoptons une approche méthodique pour appliquer et évaluer plusieurs modèles de régression quantile. L'objectif est de sélectionner le modèle le plus performant, qui sera ensuite utilisé dans la troisième partie du projet, dédiée à la prédiction conforme appliquée à la régression.

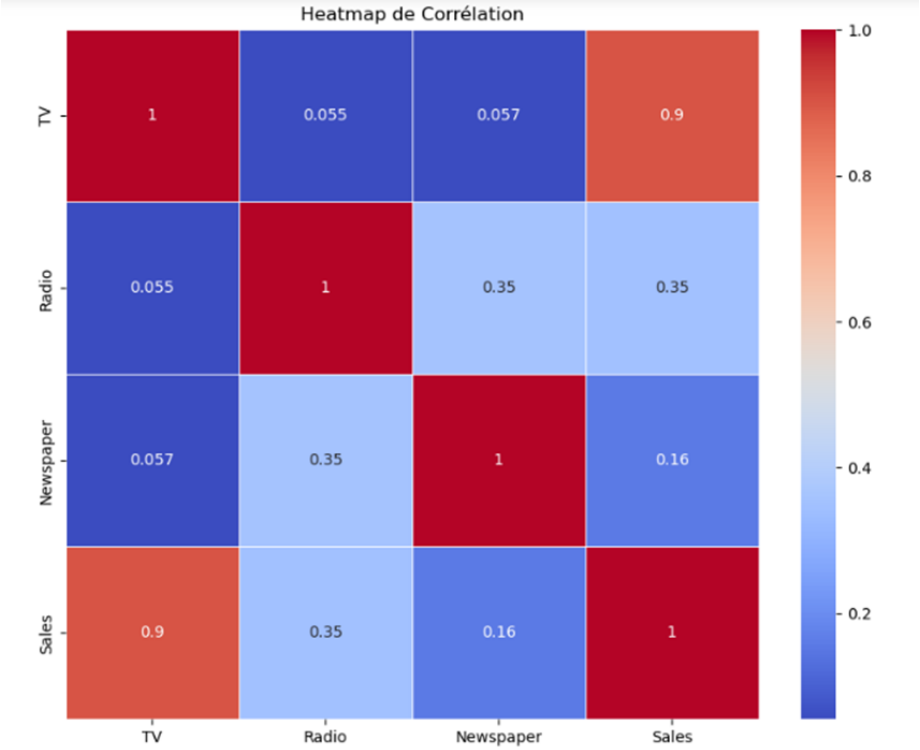
2.1 Choix du Jeu de données

Le jeu de données "Advertising", qui comprend 200 enregistrements de dépenses publicitaires dans les médias TV, Radio et Journaux, ainsi que les ventes correspondantes, est un choix judicieux pour examiner l'influence de différents types de publicités sur les performances de vente. Chaque variable du dataset, représentant les dépenses dans un canal publicitaire spécifique, fournit une valeur numérique indiquant le montant investi, et la variable cible, les Ventes, mesure l'efficacité de ces investissements en termes de résultats de vente. L'application de la régression quantile à ce dataset transcende les limites des analyses traditionnelles en offrant une exploration complète de la distribution des réponses de vente, examinant ainsi l'impact des dépenses publicitaires à travers toute la gamme des performances de vente. Cette méthode se distingue particulièrement dans la détection des effets des dépenses publicitaires sur les segments extrêmes de la distribution des ventes et se révèle robuste face aux valeurs aberrantes, un atout essentiel pour des estimations fiables dans le domaine de la publicité où des variations extrêmes sont fréquentes.

2.2 Déscription des données

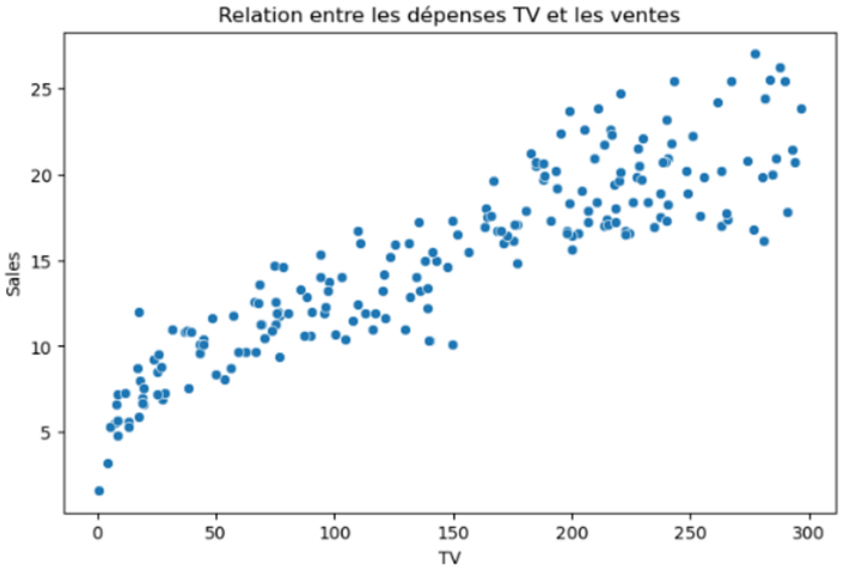
	count	mean	std	min	25%	50%	75%	max
TV	200.0	147.0425	85.854236	0.7	74.375	149.75	218.825	296.4
Radio	200.0	23.2640	14.846809	0.0	9.975	22.90	36.525	49.6
Newspaper	200.0	30.5540	21.778621	0.3	12.750	25.75	45.100	114.0
Sales	200.0	15.1305	5.283892	1.6	11.000	16.00	19.050	27.0

Les dépenses publicitaires moyennes pour la TV se distinguent nettement avec un budget moyen de 147.0425 unités monétaires, reflétant l'importance prépondérante attribuée à ce médium par les entreprises. Cette prédominance est appuyée par une médiane de 149.75, indiquant que la moitié des campagnes publicitaires dépasse ce niveau de dépense, suggérant ainsi des investissements conséquents dans ce canal. En revanche, le budget moyen consacré aux journaux est nettement inférieur avec 30.5540 unités monétaires, tandis que celui alloué à la radio, quoique supérieur à celui des journaux avec 23.2640, reste significativement moins élevé que pour la TV.
L'écart-type des dépenses en TV, estimé à environ 85.854, traduit une variabilité substantielle des investissements publicitaires dans ce médium, révélant une diversité d'approches stratégiques adoptées par différentes campagnes. Cela contraste avec les écarts-types plus modérés observés pour la radio et les journaux, respectivement de 14.846 et 21.778, qui indiquent une uniformité plus marquée dans les niveaux de dépense. Ces observations statistiques suggèrent que la TV est privilégiée comme canal de diffusion publicitaire principal.



La visualisation sous forme de heatmap de corrélation pour le dataset "Advertising" révèle des associations quantitatives entre les dépenses publicitaires sur les canaux TV, Radio et Journaux et les ventes réalisées. Une corrélation élevée de 0,9 entre les dépenses TV et les ventes indique une forte relation positive, suggérant que les investissements en publicité télévisée sont étroitement liés à l'augmentation des ventes. Cette association robuste peut signaler que la télévision, en tant que médium publicitaire, détient une influence considérable sur le comportement d'achat des consommateurs.

En ce qui concerne la radio, une corrélation de 0,35 avec les ventes, bien que modeste par rapport à la TV, démontre néanmoins une relation positive significative. Cela implique que les dépenses publicitaires en radio ont également un impact positif sur les ventes, bien que dans une moindre mesure que la publicité télévisée. Quant aux journaux, la corrélation avec les ventes est la plus faible, à 0,16, ce qui peut suggérer que bien que les dépenses publicitaires dans les journaux aient un lien positif avec les ventes, cet effet est moins prononcé que pour les médias numériques. Cela pourrait indiquer que les investissements dans la publicité imprimée sont moins efficaces pour stimuler les ventes par rapport aux options numériques, possiblement reflétant une transition des consommateurs vers les médias numériques et une portée plus limitée des journaux.

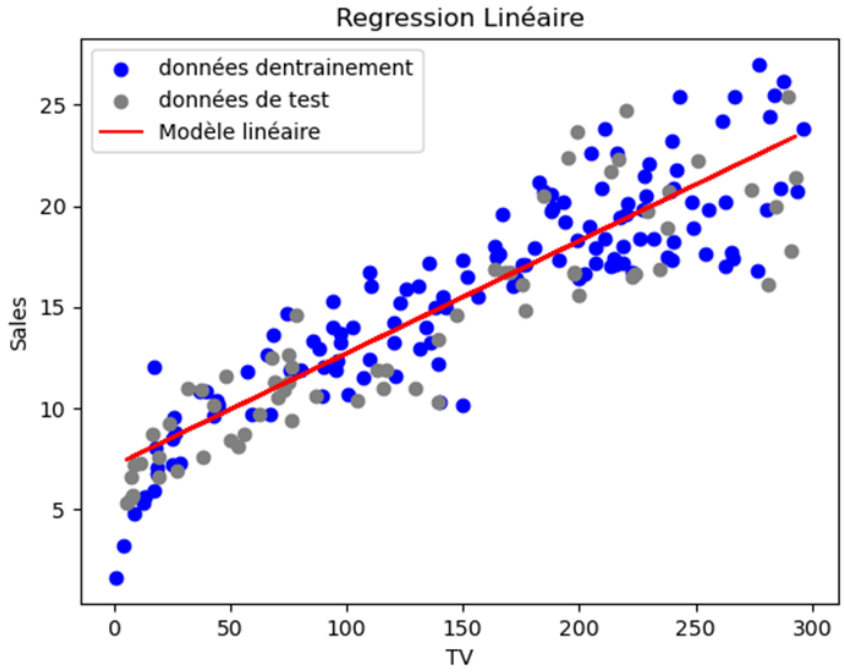


L'analyse visuelle du plot des dépenses TV en comparaison avec les ventes suggère une relation linéaire positive, où les investissements accrus en publicité télévisée semblent correspondre à une augmentation des ventes. La tendance ascendante sans la présence marquée d'outliers extrêmes indique une cohérence dans cette relation, renforcée par une légère courbure suggérant une accélération des ventes avec des dépenses publicitaires plus élevées. Cette corrélation de 0,9 entre les dépenses TV et les ventes illustre l'importance de la télévision comme canal publicitaire efficace, offrant à l'entreprise une opportunité significative d'optimiser son budget publicitaire pour maximiser le retour sur investissement.

Nous allons focaliser sur la variable TV pour notre étude et essayer de comprendre en profondeur la relation entre les dépenses publicitaire dans cette variable et les ventes (Sales).

2.3 Application de la regression linéaire

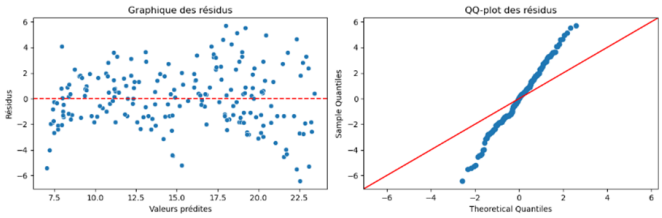
Comme nous venons de voir, il y a une tendance apparente qui montre l'existence d'une relation linéaire entre TV et Sales, ce qui justifie l'application de la régression linéaire.
La régression linéaire est une méthode statistique utilisée pour modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Fondée sur le principe des moindres carrés, cette technique cherche à établir une ligne droite (modèle linéaire) qui minimise la somme des carrés des écarts entre les valeurs observées et celles prédites par le modèle. Dans sa forme la plus simple, la régression linéaire simple, le modèle relie deux variables par une équation linéaire de la forme y=a+bx, où y est la variable dépendante, x la variable indépendante, b la pente de la ligne, et a l'ordonnée à l'origine.
Cela nous permettrait de prédire les ventes en fonction de dépenses TV spécifiques.
La pertinence du modèle de régression linéaire est habituellement vérifiée par l'analyse de MSE, et le coefficient de détermination (R2).




```
R2 pour la regression linéaire: 0.8025872552039545
MSE de la regression linéaire : 2.3559186485498715
```

Le R2 est de 0.8028, ce qui suggère que le modèle est capable d'expliquer environ 80.28% de la variance de la variable dépendante (les ventes) à partir de la variable indépendante (dépenses TV). En ce qui concerne l'erreur de prédiction, la Mean Squared Error (MSE) est de 2.3559. Compte tenu de la large variabilité des ventes, cette valeur de MSE est relativement faible, impliquant une précision de prédiction élevée du modèle. L'adéquation du modèle est également corroborée par l'observation graphique, où la ligne de régression ajustée correspond bien aux données d'entraînement (représentées en bleu) et s'aligne de manière satisfaisante avec les données de test (en gris). Cette cohérence suggère une bonne généralisation du modèle.

Pour confirmer la fiabilité du modèle, il est essentiel de vérifier les hypothèses sous-jacentes à la régression linéaire. Ces vérifications comprennent l'examen de l'homoscédasticité, de l'indépendance des résidus, de la normalité, et de l'absence de multicollinéarité, assurant ainsi la validité et l'intégrité du modèle.



Variable: TV

- Test d'homoscédasticité (Breusch-Pagan) p-value: 0.0001
- Test de normalité (Shapiro-Wilk) p-value: 0.5269
- VIF (Facteur d'Inflation de la Variance): 2.2381

Graphique des résidus :

Le graphique des résidus montre la dispersion des résidus (erreurs de prédiction) par rapport aux valeurs prédites. Dans le graphique, bien qu'il semble y avoir une dispersion aléatoire, une certaine structure pourrait être présente, suggérant que le modèle pourrait ne pas capturer entièrement la complexité des données ou qu'il y a des effets non linéaires.

QQ-plot des résidus :
Le QQ-plot (Quantile-Quantile plot) compare la distribution des résidus avec une distribution normale théorique. Dans le graphique, il y a une déviation des points de la ligne dans les extrémités, indiquant que les résidus ont des queues lourdes ou, il y a plus d'observations extrêmes que ce que la distribution normale prédirait.

Test d'homoscédasticité (Breusch-Pagan) : Avec une p-value de 0.0001, le test suggère que l'hypothèse d'homoscédasticité est rejetée, indiquant que les variances des résidus ne sont pas constantes à travers les valeurs prédites.

Test de normalité (Shapiro-Wilk) : Une p-value de 0.5269 indique qu'il n'y a pas de preuve statistique pour rejeter l'hypothèse que les résidus suivent une distribution normale.

VIF (Facteur d'Inflation de la Variance) : Un VIF de 2.2381 est en dessous du seuil couramment utilisé de 5 ou 10, suggérant que la multicollinéarité n'est probablement pas un problème pour la variable TV dans ce modèle.

Au final, bien que la variable TV ait une forte relation avec les ventes, le modèle actuel pourrait être amélioré. L'hétéroscédasticité des résidus indique que les erreurs de prédiction ne sont pas uniformes à travers toutes les valeurs prédites, ce qui pourrait être dû à des effets non modélisés ou à une relation non linéaire entre les dépenses TV et les ventes. La présence de résidus avec des queues lourdes pourrait indiquer des points de données influents ou des anomalies qui affectent les prédictions.

Cependant, l'absence de multicollinéarité indique que la variable TV fournit des informations uniques qui ne sont pas redondantes avec d'autres prédicteurs dans le modèle. Des analyses supplémentaires, telles que l'ajustement d'un modèle de régression quantile ou la transformation des variables, pourraient être envisagées pour mieux capturer la relation entre les dépenses TV et les ventes et pour traiter l'hétéroscédasticité des résidus ce qui peut améliorer la performance du modèle. Appliquons donc d'abord la regression qunatile.

2.4 Application de la régression quantile

Comme nous venons de voir, les hypothèses sous-jacentes au modèle linéaire ne reflètent pas toujours fidèlement la complexité du monde réel.

ce qui nous pousse à chercher des méthodes plus adaptatives pour quantifier l'erreur de prédiction. La régression quantile est une méthode adaptative qui ne repose pas sur les hypothèses précédentes.

L'objectif d'une régression quantile est de déterminer comment les quantiles conditionnels se comportent selon des variables explicatives. Ainsi, la régression quantile peut sortir une régression correspondante à chaque niveau voulu.

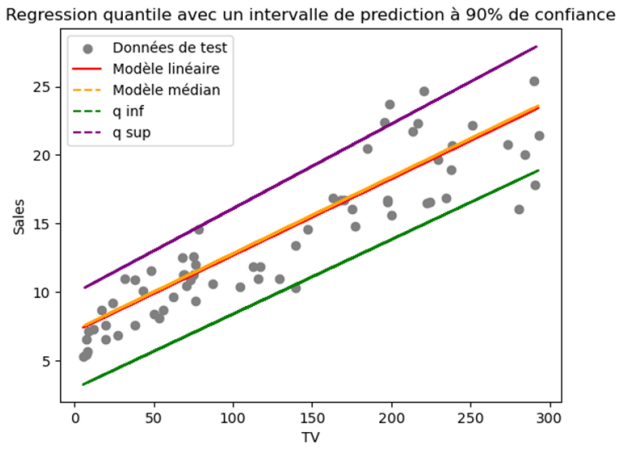
La régression quantile est particulièrement utile pour comprendre les phénomènes comme les inégalités ou les risques extrêmes, où les extrêmes de la distribution sont plus pertinents que la moyenne. Par exemple, en économie, il peut être plus instructif de comprendre comment les politiques affectent les 10% les plus pauvres ou les plus riches de la population plutôt que la population moyenne.

La régression quantile repose sur la définition du quantile d'ordre τ pour une variable aléatoire Y, exprimée par $q_\tau(Y) = F_Y^{-1}(\tau)$. Dans cette approche, la fonction quantile conditionnelle est supposée linéaire en X, donnée par $q_\tau(Y|X) = X^T\beta_\tau$. Pour estimer les coefficients, on utilise l'estimateur du quantile d'ordre τ, défini par une fonction de perte qui minimise la somme pondérée des écarts.

L'application pratique de la régression quantile implique le choix de quantiles spécifiques pour établir des intervalles de prédiction. Par exemple, pour un intervalle de prédiction de 90%, les quantiles 5% et 95% sont sélectionnés. ce qui nous permet de fournir une estimation de la plage dans laquelle les vraies valeurs des ventes sont susceptibles de se trouver pour des dépenses TV données,

avec une certaine assurance que seulement 10% des vraies valeurs tomberont en dehors de cet intervalle.

Il faut noter que un intervalle bien calibré qui capture le pourcentage attendu d'observations (par exemple 90%) indique que le modèle est bien ajusté. Si l'intervalle ne capture pas le pourcentage attendu d'observations, cela peut indiquer un problème avec le modèle ou les hypothèses sur lesquelles il repose. Ici nous avons choisi le niveau de confiance à 90% car ce niveau représente un compromis entre la précision des prédictions (assurer que la majorité des valeurs observées tombent dans cet intervalle) et la largeur de l'intervalle (ne pas le rendre trop large au point qu'il devienne peu informatif). Ce niveau de confiance élevé est choisi pour équilibrer la fiabilité des prédictions avec la nécessité de gérer l'incertitude inhérente aux données.




```
R2 pour la regression linéaire:  0.8025872552039545
MSE pour la regression linéaire: 2.3559186485498715
MSE pour la regression quantile: 2.39150074028865
```

```
Entrée [72]:
# Compter les points de test qui sont en dehors des intervalles de prédiction:
out_of_bounds = y_test[(y_test < y_pred_quantile_low) | (y_test > y_pred_quantile_high)]
number_out_of_bounds = out_of_bounds.shape[0]
print(f"Nombre de points de test hors de l'intervalle de prédiction : {number_out_of_bounds}")

Nombre de points de test hors de l'intervalle de prédiction : 6

Entrée [71]:
#compter et imprimer les couverages empiriques et théoriques:
empirical_coverage  = np.sum((y_pred_quantile_low<=y_test)*(y_test<=y_pred_quantile_high))/len(y_test)
theoretical_coverage = 1-beta
print("Theoretical coverage: ", theoretical_coverage)
print("Empirical coverage with QR: ",empirical_coverage)

Theoretical coverage:  0.9
Empirical coverage with QR:  0.9090909090909091
```

Dans le figure présenté, trois modèles de régression quantile ont été développés pour un niveau de confiance de 90

• Quantile Bas (5ème percentile) : Correspond à la borne inférieure de l'intervalle de prédiction.

• Quantile Haut (95ème percentile) : Représente la borne supérieure.

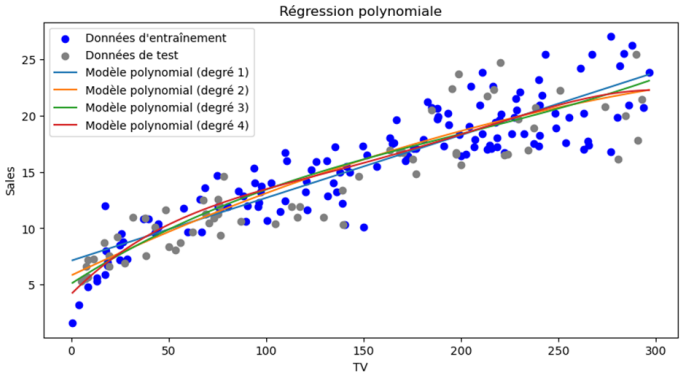
• Modèle Médian : Centre sur la médiane (50ème percentile).

La couverture empirique de l'intervalle de prédiction, mesurée à 90.99%, indique une haute précision et fiabilité, avec seulement 6 points de test en dehors de l'intervalle, conforme aux attentes théoriques.

En plus, malgré une MSE qui très proche de celle de la régression linéaire, la régression quantile est particulièrement avantageux dans notre cas (présence de l'hétéroscédasticité et des valeurs aberrantes) car la moyenne peut être influencée par les extrêmes de la distribution, tandis que la médiane et d'autres quantiles sont plus robustes à ces valeurs extrêmes. Cela fournit un cadre pour la prise de décision basée sur le risque, où une entreprise peut évaluer le potentiel de variation des ventes avec une certaine assurance.

2.5 Application de la régression polynomiale

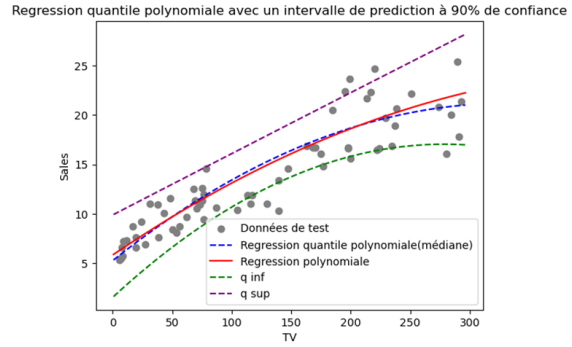
Comme nous avons vu dans le graphique du nuages de points, il semble avoir une relation non linéaire entre les dépenses publicitaires et les ventes, Si la relation n'est pas linéaire, la régression polynomiale est conçue pour capturer des relations plus complexes entre les variables, en étendant le modèle linéaire classique à des équations de degré supérieur. Nous allons tester des régression polynomiales de différents degrés et choisir celui qui capte le mieux la relation (R2 le plus élevé et MSE le plus faible).



L'analyse comparative des modèles de régression polynomiale de différents degrés indique que le modèle de degré 2 se distingue par son coefficient de détermination (R^2) élevé et sa Mean Squared Error (MSE) faible. Ce modèle surpasse la régression linéaire en termes de R^2 et MSE, et offre une meilleure performance que la régression quantile (linéaire) en termes de MSE. Ces résultats suggèrent une relation non linéaire entre les dépenses publicitaires sur la TV et les ventes. Toutefois, il convient de prendre en compte la possibilité d'un surajustement (overfitting) avec le modèle polynomial de degré 2, ce qui pourrait influencer sa performance apparente.

2.6 Application de la regression quantile de base polynomiale

Compte tenu des limitations du modèle polynomial standard, qui se concentre sur la moyenne, nous optons pour une approche de régression quantile centrée sur la médiane.. Nous appliquerons donc la régression quantile à des données transformées selon une base polynomiale de degré 2.




```
MSE de la regression polynomiale: 2.298820473278197
MSE de la regression quantile polynomiale(médiane) : 2.3228910681620447
```

```
Entrée [35]: # Compter les points de test qui sont en dehors des intervalles de prédiction:
out_of_bounds = y_test[(y_test < y_pred_quantile_low) | (y_test > y_pred_quantile_high)]
number_out_of_bounds = out_of_bounds.shape[0]
print(f"Nombre de points de test hors de l'intervalle de prédiction : {number_out_of_bounds}")

Nombre de points de test hors de l'intervalle de prédiction : 10
```

```
Entrée [36]: #compter et imprimer les couverages empiriques et théoriques:
theor_cov = 1-beta # TODO OPERAND
prop_covered_qr  = np.sum((y_pred_quantile_low<=y_test)*(y_test<=y_pred_quantile_high))/len(y_test) # TODO OPERAND
print("Theoretical coverage: ", theor_cov)
print("Empirical coverage with QR: ", prop_covered_qr)

Theoretical coverage:  0.9
Empirical coverage with QR:  0.8484848484848485
```

L'analyse de la régression quantile polynomiale révèle que, bien que son MSE soit légèrement plus élevé (de 0.02) que celui du modèle polynomial standard, sa calibration est moins idéale. Avec un R^2 légèrement supérieur à celui de la régression quantile linéaire, le modèle affiche néanmoins une couverture empirique de 84%, inférieure à la couverture théorique attendue de 90%. Cette sous-performance indique que les intervalles de prédiction ne capturent pas autant de données que prévu, suggérant une calibration imparfaite ou sur adaptation. De plus, la présence de 10 points de test hors de l'intervalle de prédiction, au lieu des 6 points maximum attendus, renforce l'idée d'une incohérence avec l'attente théorique. Ces résultats mettent en doute la capacité du modèle à fournir des prévisions fiables pour l'évaluation de la variation des ventes. Donc une entreprise ne peut pas évaluer le potentiel de variation des ventes avec une certaine assurance.

2.7 Conclusion et Choix du Meilleur Modèle

Le modèle linéaire, malgré un R^2 élevé et un MSE faible, se heurte à des limites en termes de fiabilité prédictive dues au non-respect de ses hypothèses fondamentales. Cette faiblesse est particulièrement préoccupante dans des contextes où la précision des prévisions est primordiale. L'analyse a également mis en évidence la capacité du modèle polynomial à déceler une relation plus complexe et potentiellement non linéaire entre les dépenses publicitaires et les ventes. Néanmoins, malgré ses performances supérieures en termes de R^2 et de MSE, des inquiétudes subsistent quant à sa fiabilité et au risque d'overfitting, signalées par une couverture empirique inférieure aux attentes. Par contre, la régression quantile linéaire démontre une robustesse notable. Avec un R2 de 80%, elle explique de manière significative la variabilité des ventes en fonction des dépenses publicitaires. Son MSE modéré de 2.32 traduit une erreur de prédiction raisonnable. La concordance entre la couverture empirique et théorique de 90% souligne une calibration précise et fiable, reflétant la capacité du modèle à capturer correctement les tendances des données. En somme, la régression quantile linéaire se révèle être le choix le plus judicieux, offrant un équilibre optimal entre robustesse, précision de calibration et facilité d'interprétation. Ce modèle est particulièrement adapté pour des applications pratiques où la clarté et la précision des prédictions sont cruciales.

3 Prédiction Conforme sur la Régression

3.1 Description et Approche Théorique

La prédiction conforme est une méthode statistique avancée qui fournit des prédictions accompagnées d'intervalles de confiance, offrant ainsi une mesure quantitative de la fiabilité des prédictions. Cette technique est d'une importance capitale dans les domaines où la quantification de l'incertitude est essentielle, tels que la finance, la médecine ou la recherche scientifique.

Étapes de la Prédiction Conforme:

1. Préparation et Division des Données :
• Division du jeu de données en trois sous-ensembles : train,test,calibration
2. Ajustement du Modèle de Régression sur l'ensemble d'entrainement :
• Développement d'un modèle de régression adapté sur l'ensemble d'entraînement.
3. Calcul des Scores de Conformité sur l'ensemble de calibration :
• Mesure de la conformité des nouvelles observations par rapport à celles de l'ensemble d'entraînement.

Les scores de conformité sont calculés pour chaque point de calibration (xcal,ycal) en comparaison avec une nouvelle observation xnew.

Les formules des scores de conformité sont :

- $score_{\text{plus}} = \hat{f}(x_{\text{new}}) + \left| y_{\text{cal}} - \hat{f}(x_{\text{cal}}) \right|$ (limite supérieure)
- $score_{\text{moins}} = \hat{f}(x_{\text{new}}) - \left| y_{\text{cal}} - \hat{f}(x_{\text{cal}}) \right|$ (limite inférieure)

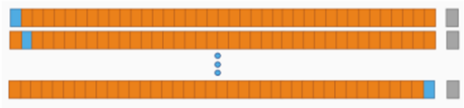
• Ces scores sont utilisés pour construire des intervalles de prédiction autour de la prédiction du modèle pour une nouvelle observation, reflétant l'incertitude estimée à partir de l'ensemble de calibration.

4. Calcul de l'Intervalle de Prédiction :
• Pour une nouvelle observation Xn+1, les bornes inférieure et supérieure de l'intervalle sont déterminées à partir des quantiles des scores de conformité de Dcal :

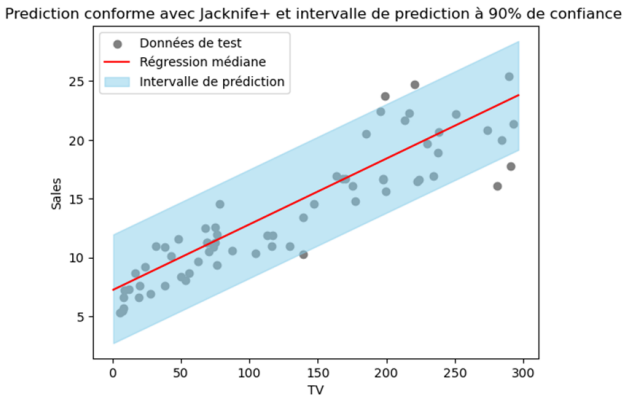
- $y_{\text{prédict}}(X_{n+1}) - \text{Quantile}(S_{\text{cal}}, 1 - \alpha/2)$
- $y_{\text{prédict}}(X_{n+1}) + \text{Quantile}(S_{\text{cal}}, \alpha/2)$
- α est le niveau de signification (par exemple, 0.1 pour un intervalle de confiance de 90%).
- L'intervalle de prédiction $\hat{Y}(X_{n+1}) \pm C(X_{n+1})$ est ainsi défini pour contenir la vraie valeur de la réponse avec une probabilité correspondant au niveau de confiance $1 - \alpha$.

3.2 Application de la méthode jacknife+

La méthode du jackknife est basé sur du leave-one-out:



Le Jackknife+ est une variante avancée de la prédiction conforme, qui emploie la technique du leave-one-out pour la calibration. Cette méthode implique la création d'un modèle pour chaque observation dans le jeu de données, en excluant temporairement cette observation spécifique. Le score de conformité est ensuite calculé pour l'observation exclue en utilisant le modèle recalibré. Ainsi, une série de scores de conformité est générée pour l'ensemble du jeu de données. Lors de la prédiction pour une nouvelle observation, ces scores sont utilisés pour établir un intervalle de prédiction.
Dans le cadre de la détermination des intervalles de prédiction en utilisant la méthode Jackknife+, le choix du niveau de confiance est crucial pour l'équilibre entre la précision et la couverture des prédictions. Un niveau de confiance élevé, tel que 95% ou 99%, élargit l'intervalle de prédiction, augmentant ainsi la probabilité d'englober la vraie valeur. Cependant, cet élargissement peut réduire la précision pour des prédictions spécifiques, en rendant l'intervalle moins informatif. Inversement, un niveau de confiance plus bas resserre l'intervalle, améliorant la précision mais au risque d'exclure la vraie valeur.
Donc, si l'objectif est de prendre des décisions à haut risque (par exemple, allouer un grand budget publicitaire), un niveau de confiance plus élevé pourrait être privilégié pour réduire le risque de prendre une décision basée sur une prédiction qui n'est pas fiable. Dans notre cas nous pouvons envisager donc un niveau de confiance à 90Cette approche sera mise en œuvre en appliquant la méthode Jackknife+ sur le modèle de régression quantile, qui a été identifié comme le meilleur modèle adapté à nos données.



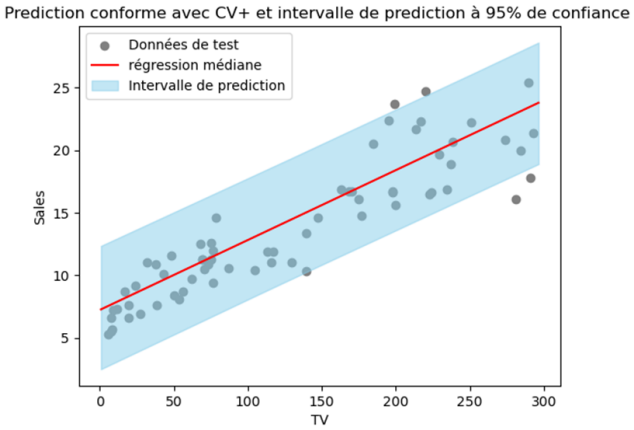
Le modèle de régression quantile démontre une capacité notable à saisir la tendance centrale et la variabilité des ventes en réponse aux variations des dépenses TV. L'intervalle de prédiction couvre efficacement la majorité des points de données de test, validant ainsi le niveau de confiance de 90% choisi pour cet ensemble de données. Cette couverture indique que l'intervalle est bien calibré et suffisamment large pour refléter l'incertitude inhérente aux données, tout en capturant la variabilité des ventes en lien avec les dépenses publicitaires. Toutefois, la présence de données extrêmes ou de valeurs aberrantes non couvertes par l'intervalle suggère la nécessité de réexaminer le modèle ou le niveau de confiance pour des applications spécifiques, en particulier là où les coûts d'erreurs de prédiction sont significatifs. En conclusion, le modèle affiche une bonne capacité prédictive, avec des intervalles de prédiction adéquats pour la variabilité observée dans les données. Les résultats sur les données de test suggèrent que les prédictions sont fiables au niveau de confiance choisi, faisant de ce modèle un outil potentiellement précieux pour guider les décisions de dépenses publicitaires TV.

3.3 Application de la méthode CV+

La prédiction conforme avec validation croisée K-Fold représente une méthode statistique avancée qui fusionne les principes de la prédiction conforme avec ceux de la validation croisée pour améliorer la robustesse et la fiabilité des prédictions dans des contextes de modélisation prédictive. Cette approche est particulièrement pertinente dans des scénarios où une estimation précise de l'incertitude associée aux prédictions est cruciale.

La méthode est basé sur le principe de cross-validation:

Le CV+, similaire au Jackknife+, utilise la validation croisée mais au lieu d'ajuster le modèle sur chaque élément dans le train set, nous allons plutôt partitionner l'ensemble d'entraînement en plusieurs sous-ensembles(k-fold)dont nous allons fixer le nombre auparavant. Le modèle est ajusté sur chaque sous-ensemble (sauf un) et les scores de conformité sont calculés pour les observations du sous-ensemble exclu. Dans le cadre de la détermination du nombre de plis (k) dans une validation croisée k-fold Un k plus petit augmente la variance mais diminue le biais, tandis qu'un k plus grand fait l'inverse. Un k trop petit pourrait ne pas capturer la diversité des données, tandis qu'un k trop grand pourrait être trop coûteux en termes de calcul et risquer d'introduire un biais. En pratique, k=5 ou k=10 sont souvent utilisés car ils offrent un bon équilibre entre efficacité de calcul et biais/variance. Nous allons fixer le nombre de k à 5.



La dispersion des points de données de test autour de la ligne de régression médiane et à l'intérieur de l'intervalle de prédiction suggère que, malgré les variations individuelles, il y a une cohérence générale dans la tendance capturée par le modèle. La largeur de l'intervalle de prédiction reflète une marge d'incertitude considérée dans les prédictions, accordant une attention à la variabilité inhérente aux données observées. Pour des décisions commerciales précises, un intervalle trop large peut ne pas fournir le niveau de détail souhaité, bien qu'il offre une assurance contre le risque de prédictions inexactes. Donc un niveau de confiance à 90% pourrait être mieux et plus précis car l'augmentation de 5% n'a pas servi à grandes choses.

3.4 Conclusion

L'adaptation de la prédiction conforme pour refléter l'hétéroscédasticité des données par l'intégration de la variance conditionnelle pourrait permettre un ajustement des intervalles de prédiction qui reflète de manière plus précise l'incertitude à différents niveaux des dépenses TV.

4 Prédiction Conforme sur la Classification

4.1 Choix du Jeu de données

Pour notre étude de classification, nous avons sélectionné un vaste ensemble de données composé de 25 000 images de scènes naturelles, chacune de taille 150x150 pixels. Cet ensemble est catégorisé en six classes distinctes représentant diverses caractéristiques géographiques et urbaines, y compris les bâtiments (0), les forêts (1), les glaciers (2), les montagnes (3), les mers (4) et les rues (5). Ces images sont préalablement divisées en ensembles spécifiques pour l'entraînement (14 000 images), le test (3 000 images) et la validation (7 000 images), chacun doté de labels numériques pour une classification systématique. Cette collection d'images offre un terrain d'application pour des défis opérationnels tels

que l'automatisation de la classification d'images, essentielle pour des initiatives variées incluant la surveillance environnementale, l'amélioration des systèmes de navigation et l'enrichissement des bases de données géographiques. L'exactitude de la classification des images selon leur contenu géographique est d'une importance capitale pour les entités engagées dans la planification urbaine et la conservation de la nature, ainsi que pour les plateformes de diffusion de contenu visuel. L'intégration de techniques de prédiction conforme dans la classification des images de scènes naturelles promet une évolution substantielle dans le déploiement de systèmes automatisés. Ces systèmes ne se contentent pas de fournir des classifications précises, mais offrent également une estimation explicite de l'incertitude liée à chaque prédiction. L'adoption de telles méthodes avancées peut renforcer la confiance dans les décisions prises de manière automatisée et optimiser les interventions humaines, contribuant ainsi de manière significative à des applications concrètes allant de la gestion des écosystèmes à la planification spatiale urbaine et rurale.

4.2 Déscription et approche théorique

La méthode de Split Conformal Prediction (SCP) est une approche rigoureuse de prédiction avec estimation de l'incertitude, qui subdivise l'ensemble de données en deux segments: un pour l'entraînement du modèle de classification et l'autre pour la calibration. Après l'entraînement sur la première partie, la SCP évalue l'atypicité de chaque observation de l'ensemble de calibration à travers une fonction de non-conformité, qui mesure la discordance avec les prédictions basées sur les données d'entraînement. Les scores de non-conformité sont ensuite utilisés pour établir les seuils prédictifs pour de nouvelles observations. Le principal avantage de la SCP est qu'elle assure, sous l'hypothèse d'indépendance et de distribution identique des données (i.i.d.), que la fréquence des erreurs de classification ne dépassera pas un niveau de signification prédéterminé ϵ sur de nouvelles données. Cette technique est robuste et agnostique quant au choix du modèle de classification, s'appliquant aussi bien aux régressions logistiques, aux Support Vector Machines (SVM), aux réseaux de neurones, et au-delà. La SCP est particulièrement précieuse dans des applications où il est crucial d'évaluer le risque associé à chaque prédiction. De plus, sa simplicité conceptuelle et sa transparence facilitent son adoption et sa compréhension, y compris pour des parties prenantes non spécialisées en statistiques. Il convient toutefois de noter que pour des données comportant des structures complexes ou des dépendances, comme les séries temporelles ou les données spatiales, l'approche SCP standard pourrait nécessiter des ajustements pour rester appropriée.

Soit X l'espace des caractéristiques et Y l'espace des étiquettes. L'ensemble d'entraînement est défini par $\{(x_i, y_i)\}_{i=1}^{n}$ et l'ensemble de calibration par $\{(x_j, y_j)\}_{j=1}^{m}$.

Un modèle de classification f est ajusté sur l'ensemble d'entraînement et attribue des scores ou probabilités aux classes pour les observations.

Fonction de Non-Conformité: La fonction de non-conformité α est définie pour quantifier l'écart des prédictions : $\alpha(x, y) = 1 - f_y(x)$, où $f_y(x)$ est la probabilité estimée que l'observation x appartienne à la classe y.

Seuils de Non-Conformité: Pour un niveau de confiance désiré $1 - \varepsilon$, le seuil t est le quantile correspondant aux scores de non-conformité de l'ensemble de calibration.

Intervalle de Prédiction: Pour une nouvelle observation x_{new}, l'ensemble de prédiction Γ est constitué de toutes les classes y pour lesquelles la non-conformité est inférieure ou égale au seuil : $\Gamma(x_{\text{new}}) = \{y : \alpha(x_{\text{new}}, y) \leq t\}$.

En résumé, la SCP offre un cadre robuste et flexible pour la classification, enrichissant les prédictions d'une dimension d'incertitude calculée, ce qui est essentiel pour une prise de décision éclairée dans des domaines exigeant une haute précision et fiabilité prédictive.

4.3 le choix de modele

Nous avons choisie le modele ResNet18, c'est un réseau de neurones convolutif connu pour son efficacité en classification d'images, est choisi pour sa capacité à traiter les problèmes de disparition du gradient grâce aux connexions résiduelles. Sa structure de 18 couches, alliant convolutions et connexions directes, optimise l'apprentissage en profondeur.

Pré-entraîné sur ImageNet, ResNet18 est idéal pour le transfert d'apprentissage, offrant de bonnes performances même avec des ressources de calcul limitées ou des données d'entraînement restreintes, ce qui en fait un choix robuste pour diverses applications de vision par ordinateur.

4.4 application de la split conformal prediction

4.4.1 Score du modele

```
Train Epoch: 1 [0/14034 (0%)]   Loss: 2.018703
Train Epoch: 1 [3200/14034 (23%)]       Loss: 0.441220
Train Epoch: 1 [6400/14034 (46%)]       Loss: 0.306337
Train Epoch: 1 [9600/14034 (68%)]       Loss: 0.425510
Train Epoch: 1 [12800/14034 (91%)]      Loss: 0.476770

Test set: Average loss: 0.0095, Accuracy: 2677/3000 (89%)
```

L'entraînement du modèle montre une réduction initiale rapide de la perte, se stabilisant vers la fin de la première époque. Sur l'ensemble de test, le modèle affiche une précision de 89%, avec une perte moyenne de 0.0895, indiquant une performance prédictive élevée dès le début de l'entraînement.

4.4.2 Resultat de la prediction conforme

les prédictions fournies par le modèle ResNet18, après application de la fonction softmax pour obtenir les probabilités de classe, et en utilisant la Split Conformal Prediction avec un seuil alpha de 0.1, indiquent un niveau de confiance de 90% dans la classification d'images de notre jeu de données. Ce processus génère des estimations de probabilité précises pour chaque classe, illustrant la forte confiance du modèle dans ses prédictions, comme en témoigne la probabilité de 98% attribuée à la classe 4 pour la deuxième image. En atteignant un seuil de confiance cumulatif de 0.99, le modèle démontre une prédiction certifiée avec un degré élevé de précision, s'assurant que les prédictions incluent la classe réelle avec une assurance de 90%. Cette fiabilité est essentielle pour des applications où des décisions critiques dépendent de l'exactitude de la classification d'images. Toutefois, une vigilance s'impose face aux valeurs aberrantes ou aux incertitudes résiduelles, qui peuvent exiger une analyse plus poussée ou un ajustement du seuil de confiance pour une assurance complète de la classification.

4.5 Conclusion

En conclusion, l'application de la Split Conformal Prediction en conjonction avec le modèle ResNet18 représente une avancée significative pour des prédictions fiables en vision par ordinateur. Cette méthodologie non seulement renforce la confiance dans la classification des images mais fournit également une estimation explicite de l'incertitude, ce qui est crucial dans des domaines nécessitant des décisions précises basées sur des analyses visuelles

Liens des bases de donnes

- Dataset1: Advertising
- Dataset2: intel_image