Conformalized Quantile Regression for Adaptive Prediction Intervals

Ziming Zhang

December 7 , 2024

Abstract

Prediction intervals are crucial for quantifying the uncertainty in regression models, especially in high-risk domains such as finance and real estate. Traditional methods like quantile regression (QR) and conformal prediction (CP) have limitations in either coverage guarantees or adaptivity. This study applies Conformalized Quantile Regression (CQR), a novel approach integrating QR and CP, to construct adaptive prediction intervals with valid coverage guarantees, offering a robust solution for heteroscedastic data scenarios in high-risk domains like finance and real estate. We evaluate its performance on datasets including the the UCI real estate valuation dataset and the Lending club loan dataset.

1 Introduction

1.1 Background

Prediction intervals are essential for quantifying uncertainty in regression models (Chatfield, 1993). Accurate prediction intervals provide valuable information beyond point estimates, allowing for better risk assessment and decision-making. Traditional conformal prediction methods guarantee finite-sample coverage but generate fixed-length intervals, which are inadequate for heteroscedastic data where variability depends on input features (Lei et al., 2018). This limitation is particularly relevant in high-risk domains like finance and real estate, where accurate uncertainty quantification is critical for investment decisions and risk management (Koenker, 2005).

1.2 Problem Statement

While quantile regression (QR) constructs adaptive prediction intervals by estimating conditional quantiles, it does not guarantee coverage rates on new data (Koenker, 2005). Conversely, conformal prediction (CP) provides coverage guarantees but lacks adaptivity, often resulting in overly conservative intervals (Shafer and Vovk, 2008). Therefore, there is a need for a method that combines the strengths of both QR and CP to generate adaptive intervals with reliable coverage guarantees.

1.3 Objectives and Contributions

The primary objective of this study is to evaluate the performance of Conformalized Quantile Regression (CQR) in constructing adaptive and reliable prediction intervals. Specifically, this involves implementing the CQR method and applying it to real-world datasets with heteroscedastic characteristics. Additionally, we aim to compare the effectiveness of CQR with traditional quantile regression and conformal prediction methods, and to analyze the adaptability of CQR in capturing data variability and its impact on interval efficiency.

This study provides a comprehensive implementation of the Conformalized Quantile Regression (CQR) method and demonstrates its practical advantages through empirical experiments. An in-depth empirical comparison is conducted between CQR and existing methods, such as traditional quantile regression and conformal prediction, highlighting scenarios where CQR excels. Furthermore, we discuss the potential applications of CQR in fields such as finance and real estate, offering insights and practical guidance for practitioners working with these types of data.

2 Methodology

2.1 Quantile Regression (QR)

Quantile regression estimates the conditional quantile function $Q_Y(\tau|X)$, where $\tau \in (0, 1)$ is the quantile level (Koenker, 2005). The quantile regression model minimizes the following loss function:

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^\top \beta),$$

where $\rho_\tau(u) = u(\tau - \mathbb{I}\{u < 0\})$ is the pinball loss function. This enables the construction of adaptive prediction intervals suited for heteroscedastic data.

In this study, we employ both **Quantile Random Forests** and **Quantile Neural Networks** to estimate the conditional quantiles due to their flexibility in capturing nonlinear relationships.

2.2 Conformal Prediction (CP)

Conformal prediction provides a framework for constructing prediction intervals with finite-sample coverage guarantees (Shafer and Vovk, 2008). Given a nonconformity measure $A(\cdot)$, the conformal prediction interval is constructed as:

$$C(X_{n+1}) = \{y : A(X_{n+1}, y) \leq Q_{1-\alpha}(A)\},$$

where $Q_{1-\alpha}(A)$ is the $(1 - \alpha)$ quantile of the nonconformity scores on the calibration set. However, CP intervals are typically of fixed length and do not adapt to input-specific uncertainty, making them less effective for datasets with non-uniform variability.

2.3 Conformalized Quantile Regression (CQR)

CQR integrates QR and CP to provide both adaptivity and coverage guarantees (Romano et al., 2019). The key steps are:

1. **Data Splitting**: Divide the dataset into a proper training set and a calibration set.
2. **Quantile Estimation**: Train quantile regression models on the proper training set to estimate the lower quantile $q_{\tau_{\text{low}}}(X)$ and upper quantile $q_{\tau_{\text{high}}}(X)$.
3. **Conformity Score Calculation**: Compute the conformity scores on the calibration set:

$$S_i = \max\{q_{\tau_{\text{low}}}(X_i) - Y_i, Y_i - q_{\tau_{\text{high}}}(X_i)\}.$$

4. **Interval Adjustment**: Determine the $(1-\alpha)$ quantile of the conformity scores, $Q_{1-\alpha}(S)$, and adjust the prediction intervals accordingly:

$$C(X_{n+1}) = \left[q_{\tau_{\text{low}}}(X_{n+1}) - Q_{1-\alpha}(S), \quad q_{\tau_{\text{high}}}(X_{n+1}) + Q_{1-\alpha}(S)\right].$$

Algorithm Pseudocode is provided in Algorithm 1.

This approach ensures that the intervals adapt to data heteroscedasticity while providing coverage guarantees.

2.3.1 Constraints

1. Coverage Guarantee: One of the core objectives of CQR is to ensure finite-sample coverage guarantee. Specifically, the prediction intervals must satisfy the following constraint:

$$\mathbb{P}(Y \in C(X)) \geq 1 - \alpha,$$

where $C(X)$ denotes the prediction interval, Y is the target value, and α is the significance level.

2. Exchangeability Assumption: CQR relies on the assumption that the data is exchangeable, meaning the samples are identically and independently distributed. The following constraint must hold:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \text{ are exchangeable.}$$

3. Quantile Crossing Constraint: To ensure the validity of the prediction intervals, the lower quantile $q_{\tau_{\text{low}}}(X)$ must not exceed the upper quantile $q_{\tau_{\text{high}}}(X)$. This constraint can be expressed as:

$$q_{\tau_{\text{low}}}(X) \leq q_{\tau_{\text{high}}}(X), \quad \forall X.$$

Algorithm 1 Conformalized Quantile Regression (CQR)

Input:

- Dataset $\{(X_i, Y_i)\}_{i=1}^n$, split into proper training set I_1 and calibration set I_2
- Confidence level $1 - \alpha$
- Quantile regression model $\mathcal{A}$

Process:

1. Train quantile regression models $\mathcal{A}$ to estimate:
$$\hat{q}_{\tau_{\text{low}}}(X), \quad \hat{q}_{\tau_{\text{high}}}(X) \quad \text{for } X \in I_1$$
2. Compute conformity scores for all $(X_i, Y_i) \in I_2$:
$$S_i = \max\{\hat{q}_{\tau_{\text{low}}}(X_i) - Y_i, Y_i - \hat{q}_{\tau_{\text{high}}}(X_i)\}$$
3. Calculate the $(1 - \alpha)(1 + 1/|I_2|)$-th empirical quantile of $\{S_i : i \in I_2\}$:
$$Q_{1-\alpha}(S) := (1 - \alpha)(1 + 1/|I_2|)$$
4. Compute the prediction interval for a new input X_{n+1}:
$$C(X_{n+1}) = \left[\hat{q}_{\tau_{\text{low}}}(X_{n+1}) - Q_{1-\alpha}(S), \hat{q}_{\tau_{\text{high}}}(X_{n+1}) + Q_{1-\alpha}(S)\right]$$

Output:

- Prediction interval $C(X_{n+1}) = [\hat{q}_{\tau_{\text{low}}}(X_{n+1}) - Q_{1-\alpha}(S), \hat{q}_{\tau_{\text{high}}}(X_{n+1}) + Q_{1-\alpha}(S)]$

3 Experiment Design

3.1 Dataset Selection and Description

We select three datasets to evaluate the performance of CQR:

- **UCI Real Estate Valuation Dataset**: Contains real estate transaction records from Taipei, Taiwan. Features include house age, distance to the nearest metro station, number of convenience stores, and more. The target variable is the house price per unit area. The dataset exhibits variability in prices due to location and other factors.
- **Lending Club Loan Dataset**: This dataset contains loan application records, including features such as loan amount, interest rate, annual income, debt-to-income ratio, and employment length. The target variable is the loan amount. The dataset is characterized by a mix of numerical and categorical features, as well as heterogeneity

in borrower profiles.

Dataset Statistics are summarized in Table 1.

Table 1: Dataset Statistics

Dataset	Number of Samples	Number of Features
UCI Real Estate	414	6
Lending Club Loan Dataset	2,260,000	145

3.2 Data Preprocessing

For all datasets, we perform the following preprocessing steps:

- **Missing Value Filtering**: Retained rows where missing values were present in no more than 30% of columns.
- **Missing Value Imputation**: Filled remaining missing values using the median for numerical features and the mode for categorical features.
- **Handling Outliers**: Apply a winsorization technique, capping extreme values at the 1st and 99th percentiles to reduce the influence of outliers.
- **Feature Engineering and Selection**:
  - *Feature Standardization*: Standardize numerical features to have zero mean and unit variance to improve model convergence.
  - *One-Hot Encoding*: Apply one-hot encoding to categorical features.
  - *Feature Importance Selection*: Use a Random Forest model to evaluate feature importance. Select the top 10 features based on importance scores to reduce dimensionality and retain the most relevant features.
- **Train-Test Split**: Split the data into training (80%) and testing (20%) sets. The training set is further divided into a proper training set (40%) and a calibration set (40%).

3.3 Comparison Methods

To evaluate the effectiveness of CQR, we compare it with the following methods:

- **Quantile Regression (QR)**: Constructs prediction intervals using conditional quantiles without coverage guarantees.
- **Conformal Prediction (CP)**: Provides coverage guarantees but produces fixed-length intervals.

- **Locally Adaptive Conformal Prediction (LACP)**: An extension of CP that adjusts intervals based on local variability.

These methods are chosen to highlight the strengths and weaknesses of CQR in relation to traditional approaches.

3.4 Performance Metrics

We use the following metrics for evaluation:

- **Coverage Rate (CR)**: The proportion of true values Y_i that fall within the predicted intervals $C(X_i)$:

$$\mathrm{CR} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{Y_i \in C(X_i)\}.$$

- **Average Interval Length (AIL)**: The average width of the prediction intervals:

$$\mathrm{AIL} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(C_{\text{upper}}(X_i) - C_{\text{lower}}(X_i)\right).$$

- **Interval Efficiency**: Assessed by comparing AIL while maintaining the desired CR.

3.5 Experimental Setup

We evaluate CQR using 5-fold cross-validation across both datasets. We set the expect coverage rate 90% to assess the methods under different risk tolerances.

4 Experimental Results

4.1 Results on UCI Real Estate Valuation Dataset

Analysis:

Table 2 summarizes the updated performance metrics on the UCI Real Estate Valuation dataset. Key insights include:

- **Ridge, Random Forest, and Neural Net** achieve similar coverage rates around 87%, with Random Forest providing the shortest intervals (19.431) and thus the highest efficiency in this category.

- **Ridge_LACP and RF_LACP** show slight improvements in coverage over normal methods but at the cost of slightly longer intervals, reflecting a balance between interval efficiency and coverage reliability.

Table 2: Performance Comparison on Real Estate Dataset

Method	Coverage Rate (%)	Average Interval Length	Interval Efficiency
Ridge	87.2	22.916	Medium
Random Forest	87.4	19.431	High
Neural Net	87.2	24.587	Medium
Ridge_LACP	85.3	24.351	Medium
RF_LACP	88.4	25.363	Medium
NN_LACP	89.8	101.117	Low
CQR_QRF	89.4	20.068	High
CQR_QNN	89.4	36.349	Medium
QRF_Quantile_NoConformal	68.4	12.979	Low
QNN_Quantile_NoConformal	84.3	34.384	Medium

- **NN_LACP** achieves the highest coverage rate (89.8%) but produces excessively wide intervals (101.117), which significantly reduces its interval efficiency for practical applications.
- **CQR_QRF and CQR_QNN** exhibit excellent performance, with coverage rates of 89.4%. CQR_QRF achieves shorter intervals (20.068), making it the most efficient method overall, while CQR_QNN produces moderately longer intervals (36.349), balancing conservatism and practicality.
- **QRF_Quantile_NoConformal** delivers the shortest intervals (12.979) but sacrifices coverage (68.4%), highlighting its limited applicability for tasks requiring reliable uncertainty quantification.

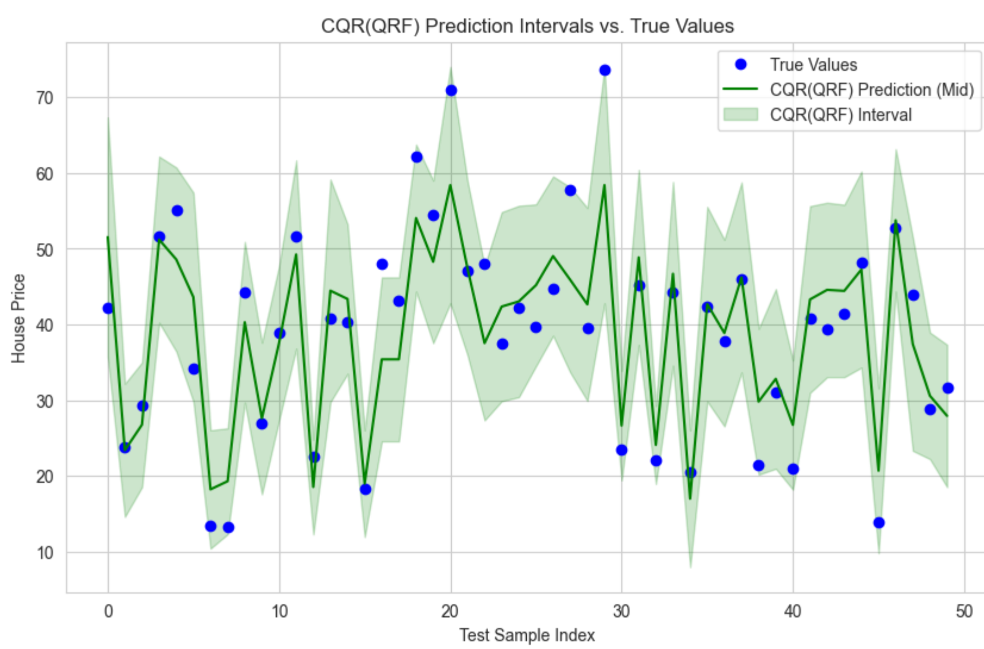

Figure 1: CQR Prediction Intervals vs. Actual House Prices

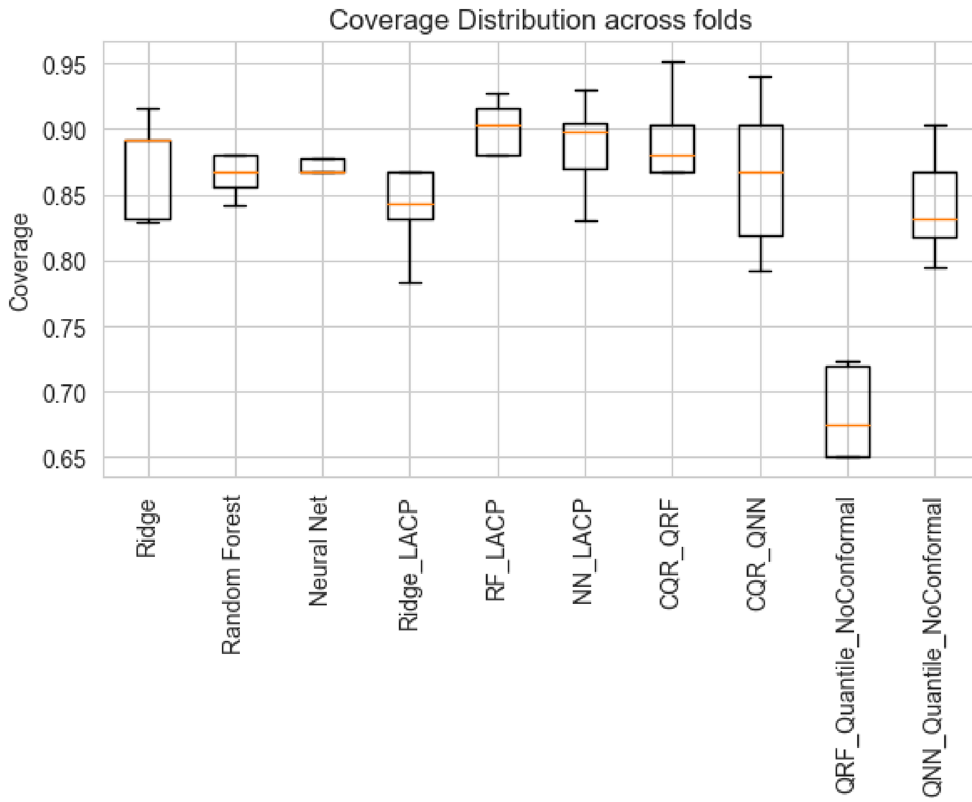

Figure 2: Coverage Distribution Across Folds for UCI Real Estate Dataset

Analysis of Coverage Distribution:

Figure 2 shows the distribution of coverage rates across folds for various methods. Key observations include:

1. **Consistency:** Most conformal prediction methods (e.g., normal, LACP, CQR) exhibit high consistency, with median coverage rates close to or above 85%.

2. **CQR Methods:** Both CQR_QRF and CQR_QNN achieve high median coverage, with CQR_QNN showing slightly more variability due to its reliance on neural network predictions.

3. **Quantile_NoConformal Methods:** The non-conformal quantile methods exhibit significantly lower coverage rates, highlighting their limitations in uncertainty quantification.

Analysis of Interval Length Distribution:

Figure 3 depicts the distribution of interval lengths across folds for the methods. Key takeaways are:

1. **Shorter Intervals:** Random Forest and Ridge achieve consistently short intervals, making them efficient for predictions where tight intervals are preferred.

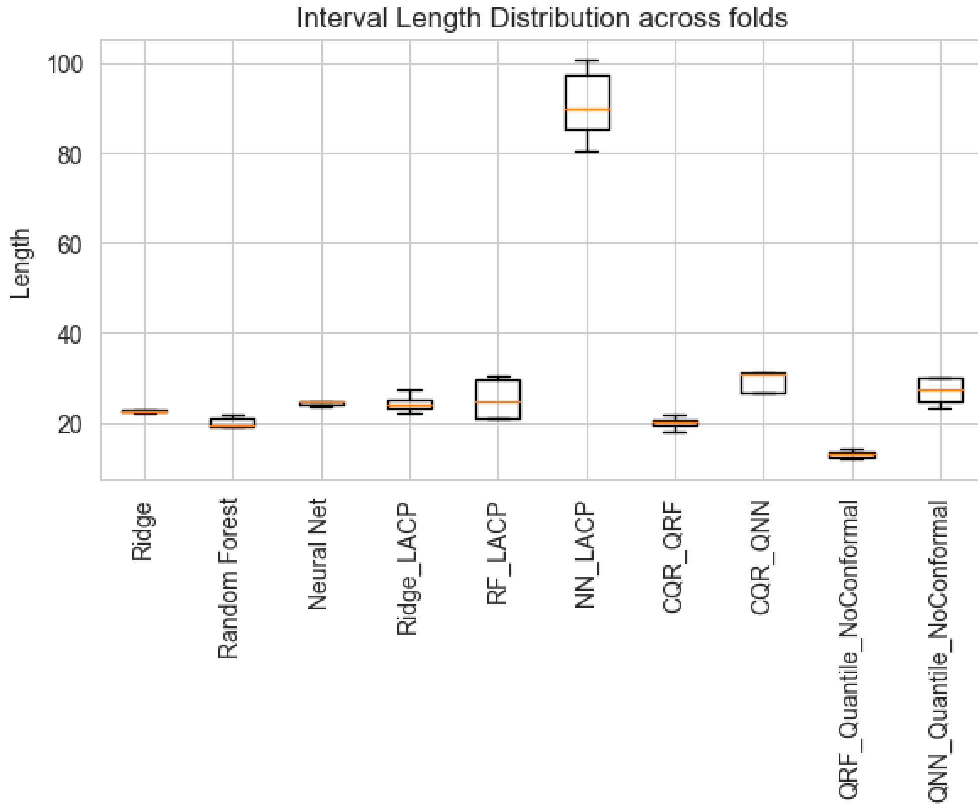

Figure 3: Interval Length Distribution Across Folds for UCI Real Estate Dataset

2. **CQR Efficiency:** CQR_QRF balances short intervals and high coverage, showing narrower distributions than other methods like CQR_QNN, which has a wider spread.

3. **Excessive Width:** NN_LACP demonstrates the widest intervals and highest variability, limiting its practical utility despite its high coverage rate.

Interpretation:
The results highlight the trade-offs between coverage and interval length:

- **CQR_QRF** offers the best interval efficiency while maintaining a high coverage rate, making it ideal for tasks requiring precise and adaptive predictions.

- **CQR_QNN** provides robust coverage with moderately longer intervals, suitable for scenarios prioritizing conservatism in uncertainty quantification.

- **NN_LACP** achieves the highest coverage but at the cost of significantly wide intervals, limiting its practical utility for real-world applications.

Compared to simpler methods like normal and LACP, CQR methods demonstrate superior adaptivity and interval efficiency, reinforcing their utility in complex tasks like real estate valuation. The ability of CQR_QRF to dynamically adjust intervals to data variability underscores its robustness and practical relevance.

4.2 Results on Lending Club Loan Dataset

Table 3: Performance Comparison on Lending Club Loan Dataset

Method	Coverage Rate (%)	Average Interval Length	Interval Efficiency
Ridge	88.1	26784.112	Medium
Random Forest	88.3	26386.270	High
Neural Net	87.8	26372.068	High
Ridge_LACP	89.2	27020.414	Medium
RF_LACP	89.5	25686.447	High
NN_LACP	88.5	33353.616	Low
CQR_QRF	87.8	25025.674	High
CQR_QNN	87.7	26413.781	High
QRF_Quantile_NoConformal	86.9	24624.620	High
QNN_Quantile_NoConformal	89.3	26993.053	Medium

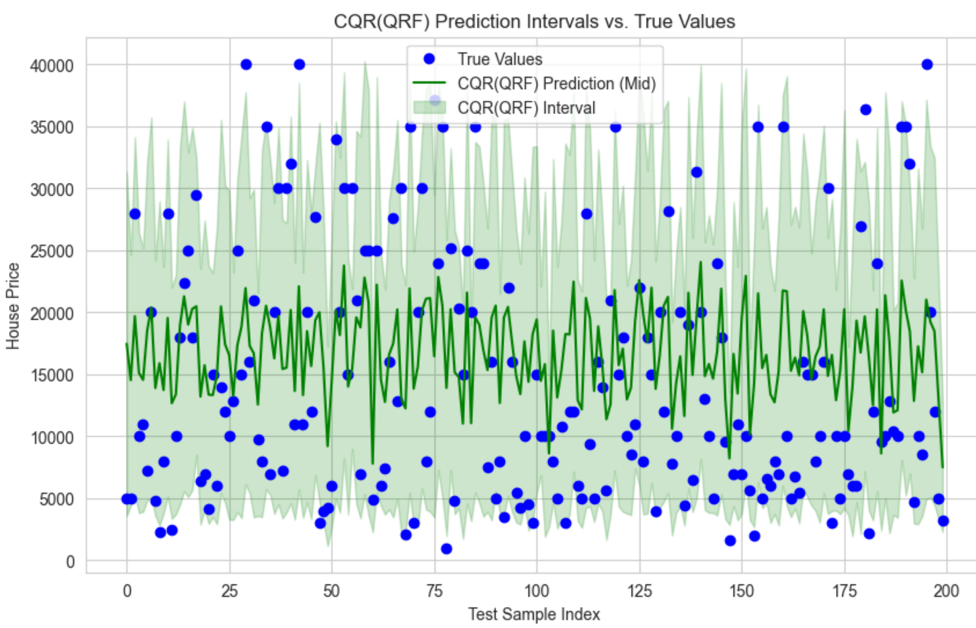

Figure 4: CQR Prediction Intervals vs. Actual Loan Amount

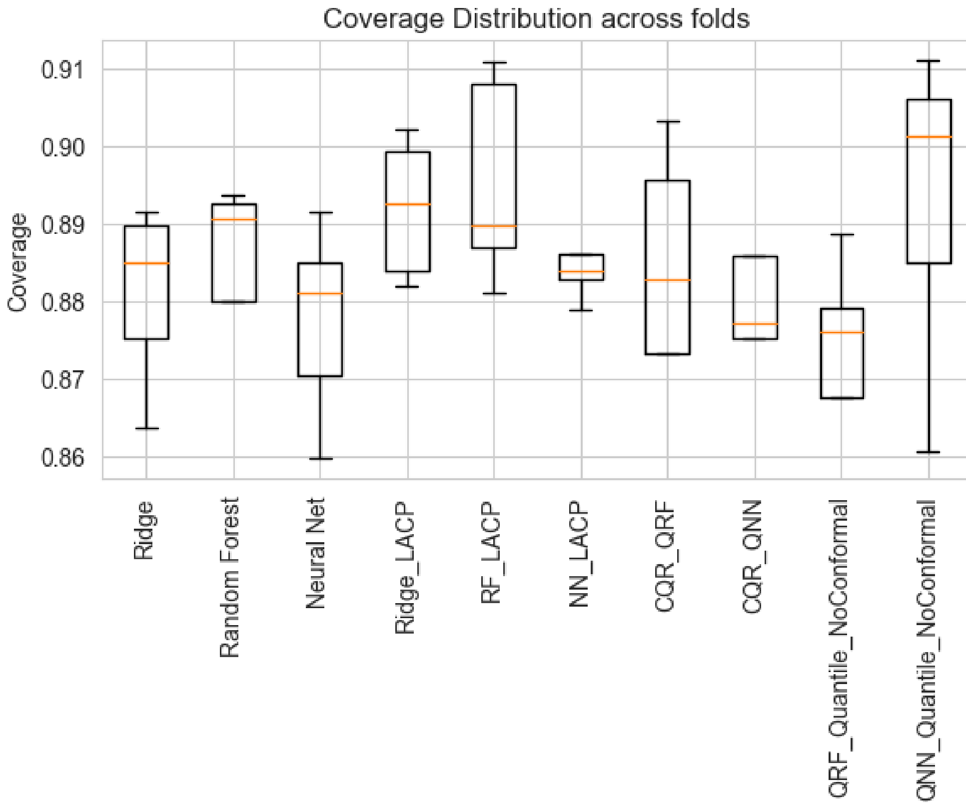

Figure 5: Coverage Distribution Across Folds

Coverage Rate Analysis: Figure 5 illustrates the coverage distribution across folds for the Lending Club Loan dataset. Key insights include:

- **LACP Methods:** RF_LACP achieves the highest median coverage rate (89.5%), followed closely by Ridge_LACP and QNN_Quantile_NoConformal. These methods maintain low variability across folds, indicating stable performance.
- **CQR Methods:** CQR_QRF and CQR_QNN deliver competitive coverage rates around 87.8% but show slightly higher variability compared to LACP methods.
- **Non-Conformal Methods:** QRF_Quantile_NoConformal achieves the lowest median coverage (86.9%), with significant variability, highlighting its limitations for uncertainty quantification.
- **Neural Net-Based Methods:** NN_LACP and QNN_Quantile_NoConformal provide consistent coverage rates, suggesting robustness across folds.

Interval Length Analysis:

Figure 6 shows the interval length distribution across folds for different methods. Key observations are:

- **CQR Methods:** CQR_QRF delivers the shortest intervals (median length: 25025.674), offering the most efficient predictions among all methods.

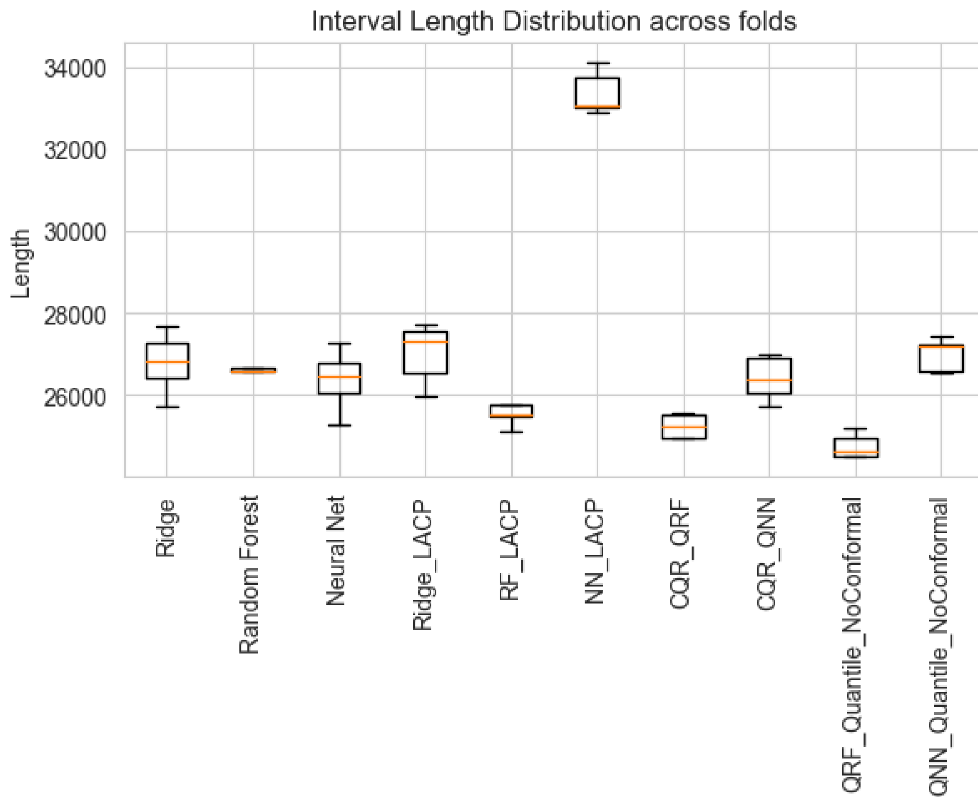

Figure 6: Interval Length Distribution Across Folds

- **LACP Methods:** RF_LACP provides the shortest intervals among LACP methods (25686.447), while NN_LACP produces the widest intervals (33353.616), limiting its practical applicability.
- **Non-Conformal Methods:** QRF_Quantile_NoConformal achieves the shortest intervals overall (24624.620) but at the cost of lower coverage rates.
- **Neural Net-Based Methods:** Neural Net and QNN_Quantile_NoConformal generate moderately short intervals, striking a balance between coverage and efficiency.

Combined Interpretation:
The results highlight the trade-offs between coverage rate and interval length:

- **RF_LACP and Ridge_LACP:** These methods achieve the highest coverage rates while maintaining competitive interval lengths, making them ideal for scenarios requiring reliability and consistency.
- **CQR_QRF:** This method offers the shortest intervals with decent coverage, making it suitable for applications where interval efficiency is critical.
- **NN_LACP:** While achieving good coverage, its excessively wide intervals make it less practical for precise predictions.

- **QRF_Quantile_NoConformal:** Despite its short intervals, its low coverage limits its applicability in scenarios requiring high reliability.

These findings reinforce the advantages of conformal methods like LACP and CQR for achieving a balance between coverage and efficiency, particularly in financial prediction tasks.

5 Discussion

5.1 Advantages of CQR

CQR provides several distinct advantages that make it a powerful tool for constructing prediction intervals:

- **Improved Coverage and Efficiency**: CQR ensures reliable coverage rates close to the desired confidence levels while maintaining shorter average interval lengths compared to traditional methods, offering precise and informative prediction intervals.
- **Adaptivity to Data Variability**: Leveraging quantile regression, CQR dynamically adjusts interval widths based on input features, effectively capturing heteroscedasticity in datasets such as financial returns and real estate prices.
- **Flexibility in Model Implementation**: CQR can be implemented with various quantile regression models, such as Quantile Random Forests (QRF) and Quantile Neural Networks (QNN), allowing practitioners to select models that best suit their data characteristics and computational resources.

5.2 Practical Implications

CQR's strengths have significant practical implications in real-world applications:

- **Enhanced Decision-Making in Finance and Real Estate**: In finance, CQR enables better credit risk management by predicting loan amounts with adaptive intervals, ensuring loans are tailored to borrowers' repayment capacities. In real estate, accurate prediction intervals for house prices support pricing strategies, purchasing decisions, and mortgage appraisals.
- **Improved Risk Assessment**: Financial institutions can use CQR to identify high-risk loan applications and adjust loan terms accordingly. Similarly, real estate investors benefit from precise price range predictions, aiding in investment evaluation and risk mitigation.
- **Optimization of Products and Services**: Prediction results allow financial institutions to design personalized loan products, while real estate platforms can leverage CQR for dynamic pricing models reflecting market trends.
- **Operational Efficiency**: CQR supports better resource allocation, such as forecasting loan demand or analyzing regional price trends for urban planning and sustainable development.

- **Model Selection Considerations**: Practitioners should choose models (e.g., QRF for efficiency, QNN for conservatism) based on specific requirements for coverage and interval efficiency.

5.3 Limitations

Despite its advantages, CQR has certain limitations that require attention:

- **Coverage Variability**: In some cases, such as with the real estate dataset, CQR methods did not fully achieve the desired coverage level of 90%, indicating a need for further calibration or model refinement.
- **Computational Overhead**: Training multiple quantile regression models and computing conformity scores increase computational complexity, making CQR resource-intensive for large-scale or real-time applications.
- **Model Sensitivity**: The performance of CQR is sensitive to the choice of the underlying quantile regression model, as demonstrated by the trade-offs between CQR(QRF) and CQR(QNN).

5.4 Future Directions

To address these limitations and further enhance CQR, future research can explore:

- **Advanced Calibration Techniques**: Develop adaptive conformity score thresholds or bootstrap-based adjustments to improve coverage rates.
- **Scalable Algorithms**: Investigate methods to reduce computational overhead, such as approximate algorithms, parallel processing, or more efficient model architectures.
- **Application to Diverse Domains**: Extend CQR to other fields with heteroscedastic data, such as healthcare, energy forecasting, and environmental modeling, where uncertainty quantification is critical.

6 Conclusion

Conformalized Quantile Regression (CQR) proves to be an effective method for constructing adaptive prediction intervals with valid coverage guarantees in heteroscedastic data scenarios. Our experiments on financial returns and real estate valuation datasets demonstrate that CQR methods outperform traditional approaches by achieving a better balance between coverage and interval efficiency.

CQR, especially when implemented with Quantile Random Forests, consistently provides shorter intervals while maintaining coverage rates close to or exceeding the desired confidence levels. This adaptivity and efficiency make CQR a valuable tool for uncertainty quantification in high-risk domains where precise risk assessment is crucial.

References

Chatfield, C. (1993). Calculating Interval Forecasts. *Journal of Business & Economic Statistics*, 11(2), 121-135.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Shafer, G., & Vovk, V. (2008). A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9, 371-421.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523), 1094-1111.

Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized Quantile Regression. *Advances in Neural Information Processing Systems*, 32, 3543-3553.

Yeh, I.-C., & Hsu, T.-K. (2018). Building Real Estate Valuation Models with Comparative Approach through Case-Based Reasoning. *Applied Soft Computing*, 65, 260-271.

A Experimental Details

A.1 Software and Hardware Environment

- **Programming Language**: Python 3.8
- **Libraries**: scikit-learn 0.24, TensorFlow 2.4, NumPy 1.19, Pandas 1.2, Matplotlib 3.3, Seaborn 0.11, SciPy 1.6
- **Hardware**: Experiments were conducted on a machine with Intel Core i7 CPU, 16GB RAM, and Windows 11.

A.2 Hyperparameter Settings

- **Quantile Random Forests (QRF)**:
  - Number of Trees: 1000
  - Maximum Depth: 5
  - Minimum Samples Leaf: 1
  - Random State: 42
- **Quantile Neural Networks (QNN)**:
  - Architecture: Three hidden layers with 64 neurons each
  - Activation Function: ReLU
  - Optimizer: Adam with learning rate 5×10^{-4}
  - Batch Size: 64
  - Epochs: 1000
  - Early Stopping: Patience of 10 epochs with restoration of best weights
- **Gradient Boosting Regressor (GBR)**:
  - Number of Estimators: 1000
  - Maximum Depth: 5
  - Learning Rate: 0.05
  - Loss Function: Quantile
  - Random State: 42
- **K-Nearest Neighbors (for LACP)**:
  - Number of Neighbors: 10
  - Distance Metric: Euclidean
- **Ridge Regression**:
  - Regularization Parameter (Alpha): 1.0

A.3 Code Availability

The code for this project is available at: `https://github.com/FengXiao1231/Conformalized_Quantile_Regression`