Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

Data Analytics for Business

MASTER'S FINAL WORK

DISSERTATION

CONFORMAL PREDICTION OF USED CAR PRICES

EDOARDO DE BESI

SUPERVISION:

João Afonso Bastos

NOVEMBER - 2023

GLOSSARY

CP conformal prediction. i, 1, 2, 4, 19

CQR conformal quantile regression. i, 2, 15, 16, 18–20

FQR frequentist quantile regression. i, 2, 15, 16, 18–20

LightGBM Light Gradient Boosting Machine. i, ii, 6, 7, 20

MAE mean absolute error. i, 12, 15, 20

ML machine learning. i, 1, 2, 10, 19

NCP nominal coverage probability. i, 2, 3, 17–20

ABSTRACT

This academic thesis addresses a critical gap in the existing literature surrounding predictive analytics and used car prices, specifically where research predominantly focuses on estimating point predictions of prices using machine learning without providing a measure of uncertainty associated with these predictions. The objective is to calculate prediction intervals using both conformal quantile regression and frequentist quantile regression on a "Light Gradient Boosting Machine (LightGBM)" model trained with a comprehensive dataset of used car listings collected in May 2021 from the United States marketplace.

The paper empirically compares these two methodologies at various nominal coverage probabilities. Notably, the study reveals a significant trade-off that decision-makers must consider—a balance between accuracy and precision. Conformal predictions uniquely offer a guarantee of the nominal coverage level at the expense of wider prediction intervals.

Furthermore, the research emphasizes that the decision on which method to use depends on the target nominal coverage probability level. As the nominal coverage probability increases, the study finds that the median width of conformal quantile regression increases more than proportionally compared to frequentist quantile regression. This implies that the coverage guarantee becomes more costly in terms of width as the nominal coverage probability rises, making conformal quantile regression more advantageous at lower nominal coverage probability.

KEYWORDS: Car Pricing; Machine Learning; Conformal Prediction; Conformal Quantile Regression.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

1 Introduction

In the current era, data has emerged as a pivotal element in decision-making, replacing traditional methods across various domains. Traditional methods of used car pricing, such as expert evaluation or the use of guidebooks, have limitations in terms of either scalability or accuracy. In rare or unusual cases, experts may offer precise valuations, however, their services might be cost-prohibitive and time-intensive, especially for straightforward vehicle evaluations. On the other hand, while guidebooks are certainly more scalable, they are often too general and are not capable of considering all factors that may influence the price of a car.

The used car sector is occupying a large share of the consumer market, with its global valuation increasing from an estimated €1.49 trillion in 2021 to €1.57 trillion in 2022 and forecasted to reach €2.57 trillion by 2030 (Grand View Research, 2022). Developing and distributing a machine learning (ML) model capable of assessing the price of used cars can have a significant global economic impact by enhancing trust, transparency, and competition within this market. Furthermore, this model can favorably impact other stakeholders of the automotive industry. For example, car rental companies with a dynamic understanding of the depreciation of their fleet can make better-informed decisions on when to sell or replace vehicles in order to maximize return of investment. Similarly, insurance companies can benefit from an accurate car valuation model by setting appropriate insurance premiums and by processing claims both with more efficiency and greater accuracy

A ML model that provides a point prediction of the price of a used car would provide a straightforward, easy to understand result that would enable immediate decision-making for all stakeholders. However, such model would lack the representation of uncertainty in price estimation and, if the price prediction is inaccurate, it could either undervalue or overvalue the car, potentially leading to financial loss or missed opportunities

Linear regression methods are capable of providing straightforward solutions that allow researchers to understand how aspects such as the production year, mileage, brand, and model influence a used car resale value. However, these methods fall short in capturing the complex non-linear relationships present in used car data (Ozgur et al., 2016). Ensemble methods address these shortcomings because, as indicated by Breiman (2001), they are capable of effectively handling complex, unstructured, and high-dimensional data. Moreover, a study by Chen et al. (2017), demonstrated that the random forest ensemble method outperforms other regression methods, especially when developing a model that is not confined to a specific car brand but includes vehicles from various brands and production years.

To the best of my knowledge, this paper expands the existing literature by employing, for the first time, conformal prediction (CP) in a used car pricing model based on an ensemble method. CP, also known as conformal inference, presents a user-friendly framework for generating statistically sound uncertainty intervals for model predictions. Moreover, CP offers clear, non-asymptotic guarantees that are not reliant on specific distributional or model assumptions, as supported by various studies including Papadopoulos et al. (2002); Vovk et al. (2005); Lei & Wasserman (2014); Angelopoulos & Bates (2023).

In Section 2, we commence with a comprehensive examination of existing research on ML models for predicting used car prices, coupled with a review of literature on CP. Section 3 is dedicated to presenting the fundamental theories and algorithms used in this study. This is followed by Section 4, where we transition from theoretical principles to their practical implementation. This segment starts by describing our dataset in Section 4.1 and then proceeds to Section 4.2, where an extensive process of data preprocessing and feature engineering is performed. Additionally, Section 4.3 offers insights into price trends and geographical distribution through visual analysis, aiding in the understanding of market dynamics and identifying any constraints within the dataset. Continuing Section 4, we delve into the construction of the model and the methodology for optimizing hyperparameters, as detailed in Section 4.4. The efficacy of the model is then evaluated, particularly by comparing the accuracy of predictive intervals generated by conformal quantile regression (CQR) and frequentist quantile regression (FQR) methods, as discussed in Section 4.5. In Section 4.6, we showcase and discuss conditional coverage across different subsets of our dataset. Lastly, in 4.7, we compare FQR and CQR as the nominal coverage probability (NCP) varies. Finally, in Section 5, we draw our conclusion and discuss the key insights derived from our research.

2 LITERATURE REVIEW

The prediction of used car prices has seen remarkable progress, evolving from early econometric models to advanced machine learning algorithms. This literature review traces the significant milestones in the field, highlighting the transition from foundational hedonic pricing methods to sophisticated contemporary approaches.

The foundational hedonic pricing models were introduced by Griliches (1961). These models break down the price of a car into its individual features—such as horsepower, weight, and brand—allowing researchers to quantify the economic value of each attribute and understand how changes in these features affect the overall price. Griliches' methodology provided a structured framework that subsequent research would build upon, enabling a more detailed analysis of car pricing.

Continuing from the foundational hedonic pricing models, the work of Berry et al. (1995), in their paper "Automobile Prices in Market Equilibrium" represents a significant milestone in the evolution of car pricing analysis. Their approach builds upon Griliches' framework by incorporating a structural model that accounts for both consumer preferences and firm behaviors in a competitive market. This model not only evaluates the impact of various car attributes on prices but also considers the strategic interactions among automobile manufacturers.

The late 20th century saw a revolution in computational power, paving the way for the integration of more sophisticated, data-driven approaches in the prediction of used car prices. This transition marked a significant shift from the static relationships defined by traditional econometric models to more dynamic methods capable of capturing complex interactions within the data. Researchers began to leverage machine learning algorithms, which could process vast amounts of data and identify intricate patterns that were previously undetectable. This evolution was driven by the need for models that could adapt to the increasing complexity and variability of the factors influencing car prices.

A pivotal study in this new era was conducted by Chen et al. (2017), who explored how different machine learning models handle varying levels of data complexity and specificity. By comparing traditional linear regression models with more advanced algorithms like Random Forests, Chen demonstrated how machine learning can better handle diverse and complex datasets. Random Forests, in particular, showed superior generalization capabilities across various car makes and models, indicating their robustness in predicting prices across a wide range of attributes and conditions. This shift emphasizes the move from static relationships observed in hedonic models to more flexible frameworks capable of capturing intricate interactions within the data.

Further expanding on this, the research by Varshitha et al. (2022) exemplifies the integration of dynamic, data-driven methods into car price prediction. Varshitha's study compared several machine learning models, including Artificial Neural Networks, Random Forests, Lasso, Ridge Regression, and traditional Linear Regression. The study found that ensemble methods, especially Random Forests, excel in managing the complex, real-world datasets used in predicting used car prices. These methods are adept at capturing a wide array of features and their interactions, offering enhanced predictive performance compared to simpler linear models.

Moreover, Han et al. (2022) focused on enhancing the predictive power of these models through sophisticated feature engineering and model optimization techniques. Han's study used correlation analysis and LightGBM for feature extraction, showing how careful selection and transformation of features could significantly improve model performance. Han further innovated by integrating Random Forest and XGBoost models into a combined regression framework, leveraging the strengths of multiple algorithms. This approach emphasized the importance of not only the choice of model but also the preprocessing and feature engineering steps preceding modeling, building upon the foundational insights provided by earlier econometric approaches.

When reviewing pertinent academic research, it is surprising that prediction intervals are largely overlooked. Models able to provide uncertainty in the form of a range of potential prices would allow stakeholders to make more informed decisions due to its ability to asses the risks inherent in the transaction. Smaller intervals imply more accurate car appraisals, while wide intervals may imply complexities in the specific use-case. Additionally, models that provide prediction intervals offer broader utility, such as in budgeting and financial planning, because they enable stakeholders to account for and prepare for potential worst-case and best-case outcomes.

In this study, in order to develop such prediction interval $C(\boldsymbol{X}_{n+1}) \subseteq \mathbb{R}$ for a specific unknown used car price Y_{n+1}, we utilize a model trained on a dataset of n explanatory variables $\{\boldsymbol{X}_i\}_{i=1}^n$ and their associated target variable price $\{Y_i\}_{i=1}^n$. Given a significance level α, we want to ensure with high probability that

$$\Pr\left(Y_{n+1} \in \mathcal{C}(\boldsymbol{X}_{n+1})\right) \geq 1 - \alpha \tag{1}$$

where $1 - \alpha$ represents the desired NCP.

Typically, the formulation of a prediction interval depends on the assumptions made about the underlying distribution of the data. It is important to note that if these assumption do not hold the reliability of the resulting prediction intervals are compromised.

An alternative method that does not rely on a distributional assumption is quantile regression. Quantile regression, as introduced by Koenker & Bassett (1978), extends the classical linear regression model by focusing on the estimation of conditional quantile functions. Unlike ordinary least squares regression, which estimates the mean of the dependent variable conditional on given covariates, quantile regression aims to estimate the α quantile of the dependent variable. The conditional quantile function is defined as follows:

$$q_{\alpha}(\boldsymbol{X}) = \inf\{Y \in \mathbb{R} : F(Y|\boldsymbol{X}) \geq \alpha\} \tag{2}$$

By training two distinct statistical models to determine the quantile functions, we can establish a predictive range for the upcoming observation, Y_{n+1}. This predictive range is referred to as the conditional prediction interval and it is bounded by the lower quantile $q_{\frac{\alpha}{2}}(X_{n+1})$ and the upper quantile $q_{1-\frac{\alpha}{2}}(X_{n+1})$.

The simulation studies and empirical evidence presented later in this paper demonstrate that, when using frequentist quantile regression, the generated prediction intervals miss the real used car prices more often than anticipated with the specified nominal coverage probability.

An efficient remedy for this issue of miscoverage of quantile regression was introduced by Romano et al. (2019). The introduced method effectively merges CP with classical quantile regression. Essentially, this innovative approach enhances CP's adaptability to heteroscedasticity, enabling it to generate prediction intervals whose lengths vary in response to the local data variability. In the context of our work, this attribute is particularly beneficial, as it equips our model to effectively manage the diverse prices and characteristics of the various cars in our dataset.

3 PREDICTIVE MODELING OF USED CAR PRICES

3.1 Methodology for Conformal Quantile Prediction and Validation

In our model, the price of a used car is the dependent variable Y while $\boldsymbol{X}$ represents the array of independent variables. Our goal is to construct a prediction interval $\mathcal{C}(\boldsymbol{X}_{n+1}) \subseteq \mathbb{R}$ for cars with given features $\boldsymbol{X}_{n+1}$, but where the market price Y_{n+1} is not known. To estimate the market value Y_{n+1}, we employ a machine learning model trained on a set of n instances $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$. When doing so, we assume that all of our observations, including the new option $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n+1}$, are exchangeable and drawn from a joint distribution $P_{X,Y}$. Our assumption is guaranteed to be valid when $(\boldsymbol{X}_i, Y_i)$ are independent and identically distributed. Within this structure, and more generally for any joint distribution $P_{X,Y}$, given a desired coverage level $1 - \alpha$, the prediction interval $\mathcal{C}(\boldsymbol{X}_{n+1})$ adheres to Equation 1.

In the process of applying conformal quantile prediction, we divide our dataset into two parts: a training subset S_1 composed of pairs $(\boldsymbol{X}_i, Y_i)$ with $i \in \mathcal{I}_1$, and a calibration subset S_2 made up of pairs $(\boldsymbol{X}_i, Y_i)$ with $i \in \mathcal{I}_2$. We then train models to estimate the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles, which we denote as

$\hat{q}_{\frac{\alpha}{2}}(\boldsymbol{X})$ and $\hat{q}_{1-\frac{\alpha}{2}}(\boldsymbol{X})$, using a quantile regression method on the training subset S_1 with the objective to achieve a coverage probability of $1-\alpha$. Next, we calculate conformity scores using the formula below:

$$\hat{\epsilon}_i = \max\left\{\hat{q}_{\frac{\alpha}{2}}(\boldsymbol{X}_i) - Y_i, Y_i - \hat{q}_{1-\frac{\alpha}{2}}(\mathbf{X}_i)\right\} \forall i \in \mathcal{I}_2. \tag{3}$$

The $1-\alpha$ quantile of the empirical distribution for these conformity scores is calculated using:

$$q_{1-\alpha}(\mathcal{I}_2) = \frac{(n_2+1)(1-\alpha)}{n_2} \text{ empirical quantile of } \hat{\epsilon}_i,\ i \in \mathcal{I}_2. \tag{4}$$

Finally, the conformalized prediction interval for the next observation Y_{n+1} is:

$$\mathcal{C}(\boldsymbol{X}_{n+1}) = \left[\hat{q}_{\frac{\alpha}{2}}(\boldsymbol{X}_{n+1}) - q_{1-\alpha}(\mathcal{I}_2), \hat{q}_{1-\frac{\alpha}{2}}(\boldsymbol{X}_{n+1}) + q_{1-\alpha}(\mathcal{I}_2)\right]. \tag{5}$$

The validity of the conformal quantile method discussed in this section is confirmed by the following theorem by Romano et al. (2019).

Theorem (Romano et al., 2019): If the pairs $(\boldsymbol{X}_i, Y_i)_{i=1}^{n+1}$ are exchangeable, then the prediction interval $\mathcal{C}(\boldsymbol{X}_{n+1})$, as defined in Equation (5), adheres to the condition:

$$\Pr(Y_{n+1} \in \mathcal{C}(\boldsymbol{X}_{n+1})) \geq 1-\alpha.$$

Furthermore, if the conformity scores $\{\hat{\varepsilon}_i : i \in \mathcal{I}_2\}$ are unique with high probability, the interval $\mathcal{C}(\mathbf{X}_{n+1})$ is nearly perfectly calibrated:

$$\Pr(Y_{n+1} \in \mathcal{C}(\boldsymbol{X}_{n+1})) \leq 1-\alpha + \frac{1}{1+n_2}.$$

3.2 *Ensemble Learning Approach with Gradient Boosting Machines*

Our approach employs an ensemble method, leveraging a variant of the gradient boosting machine to serve as the quantile regression model. This method is based on the principle of merging multiple basic models, which individually may not be very robust but collectively contribute to a more powerful and accurate prediction.

FRIEDMAN'S GRADIENT BOOST ALGORITHM

Inputs:

- dataset $(x, y)_{i=1}^{N}$
- iteration count M
- selection of the loss-function $\Psi(y, f)$
- selection of the base-learner model $h(x, \theta)$

Algorithm:

1. set $\hat{f}_0$ to a constant value

```
2. for $t = 1$ to $M$:
    (a) determine the negative gradient $g_t(x)$
    (b) fit a new base-learner function $h(x, \theta_t)$
    (c) ascertain the optimal step-size $\rho_t$ for gradient descent:
            $\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \Psi(y_i, f_{t-1}(x_i) + \rho h(x_i, \theta_t))$
    (d) update the function estimate:
            $\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t)$
3. conclude iteration
```

Examining the gradient boosting algorithm as first presented by Friedman (2001), it becomes evident that the resulting model is significantly shaped by the choices made in selecting the loss function and the base-learner models.

In our case, a series of K decision trees, denoted as $\{h_k(\boldsymbol{X})\}_{k=1}^{K}$, compose the base-learner models. We begin with a standard decision tree, $h_1(\boldsymbol{X})$, built from the original dataset. During each iteration, new trees are sequentially trained, selected, and added to our ensemble.

The new trees are trained to correct the errors produced by the ones already present in the ensemble. Then, the tree to be added is selected based on its capability to minimize the regularized loss function which, summed over all n data points, is given by:

$$\sum_{i=1}^{n} L\left(Y_i, \hat{Y}_i^{(k-1)} + h_k(\boldsymbol{X}_i)\right) + \gamma T + \frac{1}{2}\lambda \|\boldsymbol{w}_k\|^2. \tag{6}$$

More specifically, because we are predicting quantiles, the "pinball loss" function is used:

$$L\left(Y_i, \hat{Y}_i^{(k-1)} + h_k(\mathbf{X}_i)\right) = \begin{cases} \alpha\left(Y_i - \hat{Y}_i^{(k-1)} - h_k(\boldsymbol{X}_i)\right) & \text{if } Y_i \geq \hat{Y}_i^{(k-1)} + h_k(\boldsymbol{X}_i) \\ (1-\alpha)\left(\hat{Y}_i^{(k-1)} + h_k(\mathbf{X}_i) - Y_i\right) & \text{otherwise} \end{cases} \tag{7}$$

Constructing K decision trees, one after another, and then aggregating their predictions leads to the expected composite prediction of the gradient boosting machine. This cumulative prediction for $\hat{Y}$ is formulated as:

$$\hat{Y} = \sum_{k=1}^{K} h_k(\mathbf{X}). \tag{8}$$

In our research, the LightGBM algorithm is employed to effectively minimize the loss function. A distinguishing feature of LightGBM is its histogram-based learning method. Rather than evaluating every individual value for each dataset feature, LightGBM categorizes these values into discrete bins, creating histograms. This process greatly reduces memory usage and computational demands, enhancing its speed and scalability compared to other gradient boosting frameworks. Moreover, the proficiency of LightGBM in handling categorical variables and its efficient processing of datasets with numerous sparse categories further enhace its suitability for our application.

4 MODEL APPLICATION

4.1 Dataset Overview

The dataset used was downloaded from Kaggle, the author of the dataset is Austin Reese. It contains more than 425,000 listings of used cars scraped from Craigslist.com, a website owned by a private American enterprise that runs a classified ads platform. The website features various sections, including those for jobs, housing, wanted items, services, community activities, gigs, résumés, and, of course, used cars.

The listings contained in the dataset were specific to the North American market and contained cars that were built in a period between 1905 and 2022. Each row of the dataset had a specific listing posted on the website, reachable using the given URL. It is important to note that two different URLs (or rows) could be related to the same car. As for the columns of the dataset, each one represents a different feature of the car.

Column Name	Data Type	Description
id	int64	Unique identifier for each car listing.
url	object	URL link to the car listing.
region	object	Geographical region where the car is listed.
region_url	object	URL link to the specific region's page.
price	int64	Listed price of the car in USD.
year	float64	Year the car was manufactured.
manufacturer	object	Car manufacturer.
model	object	Specific model of the car.
condition	object	Condition of the car.
cylinders	object	Number of cylinders in the car's engine.
fuel	object	Type of fuel the car uses.
odometer	float64	Number of miles the car has been driven.
title_status	object	Status of the car.
transmission	object	Type of transmission in the car.
VIN	object	Vehicle Identification Number - a unique code used to identify motor vehicles.
drive	object	Drive type of the car.
size	object	Size category of the car.
type	object	Category of the car.
paint_color	object	Color of the car.
image_url	object	URL link to an image of the car.
description	object	Detailed description of the listed car.
county	float64	County information.
state	object	State where the car is listed.
lat	float64	Latitude coordinate of the car's location.
long	float64	Longitude coordinate of the car's location.
posting_date	object	Date and time when the car was listed.

TABLE I: Data Dictionary

4.2 Data Preprocessing and Feature Engineering

The accuracy of our model is related to the quality of the dataset. Therefore, before fitting our model, several changes must be performed to improve data quality.

Listings that were duplicated due to being related to the same car were identified and removed.

In order to accomplish this, the dataset was searched using the Vehicle Identification Number. This unique identifier for automobiles allowed us to pinpoint the duplicated entries that were to be removed. Subsequently, given that it does not influence the valuation of a used car, it was removed from the dataset.

Figure 1 displays a histogram based on the 'year' column of our dataset, which represents the year of manufacturing of the used car. Overlaid on the histogram, is the Kernel Density Estimation (KDE) curve, which will provide a smoothed representation of the data's distribution.

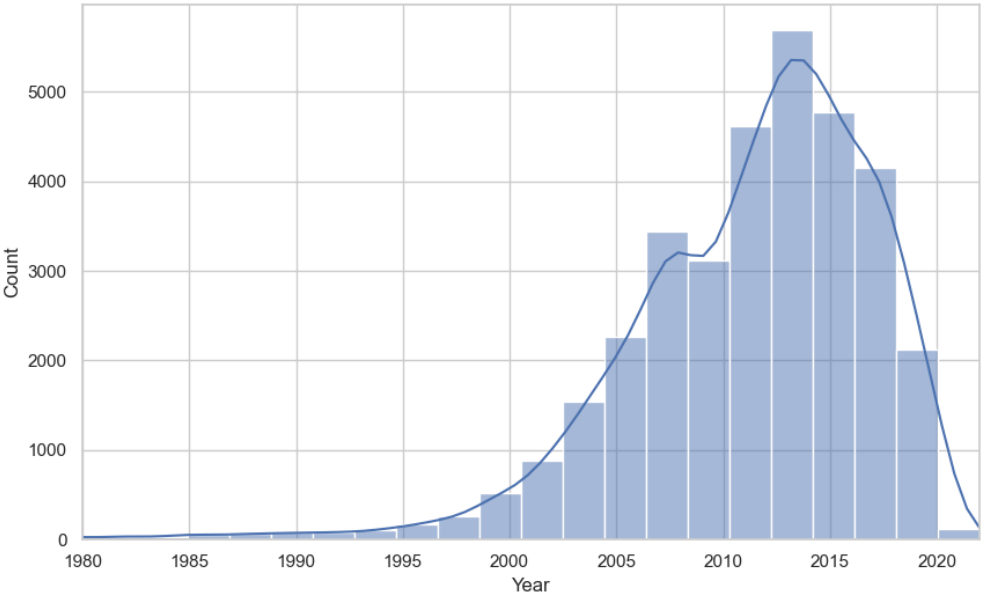


FIGURE 1: Histogram and KDE Curve Showing the Distribution of Manufacturing Years.

The above figure shows that there is a significant concentration of data between roughly 1990 and 2020. Sparse data from years outside this range could introduce noise or bias, potentially compromising the model's performance. To enhance the predictive accuracy of our model, the decision has been taken keep only the data from the period 1990-2020.

Figure 2 is a density plot which offers a visual representation of price distribution in the dataset.

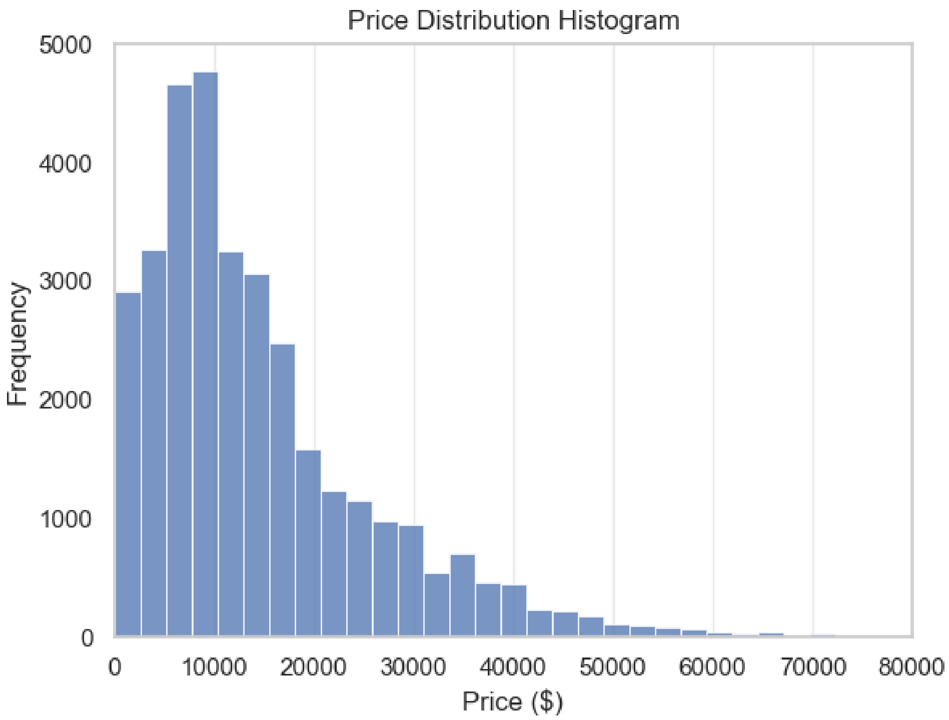


FIGURE 2: Bar Plot Displaying the Distribution of Used Car Prices in the Dataset.

Data points in the upper tail of our distribution are regarded as outliers, representing luxury vehicles. Addressing these extreme cases falls beyond the purview of this paper; hence, they have been excluded from our dataset using a max price threshold of $\$27,000$.

The data we are analyzing exhibits sparsity not only along the temporal axis but also in the categorical variables. As we can observe in Figure 3, there are certain categories that have very few instances relative to the dimension of the dataset. For example, if we look at the categorical variable 'Title Status', category 'parts only' appears only in 3 entries out of 28.057. To enhance the precision of our model, we will now systematically tackle these instances of categorical variable sparsity.

For the categorical variables 'Cylinders' and 'Fuel', our dataset is heavily skewed towards the values '4 cylinders', '6 cylinders', '8 cylinders', and 'gas' respectively. To mitigate this bias, we performed data filtering, removing entries with values other than these specified ones from our dataset.

It is out of the scope of our ML model to price used cars that are used or are not sold as 'clean'. A car with a 'clean' title is indicative of a history free from significant damage, accidents, or other factors that might unexpectedly impact its value. Consequently, listings that did not align with the 'clean' title requirement or had a 'condition' labeled as 'new' or 'salvage' were removed from the dataset.

To address the issue of sparse data in the 'fuel' feature, we leveraged domain expertise. Hybrid, electric, and other alternative fuel vehicles share certain characteristics that can be relevant for predicting used car prices. For instance, they tend to be more energy-efficient, have different maintenance requirements, and may benefit from specific government incentives or regulations. Consequently,

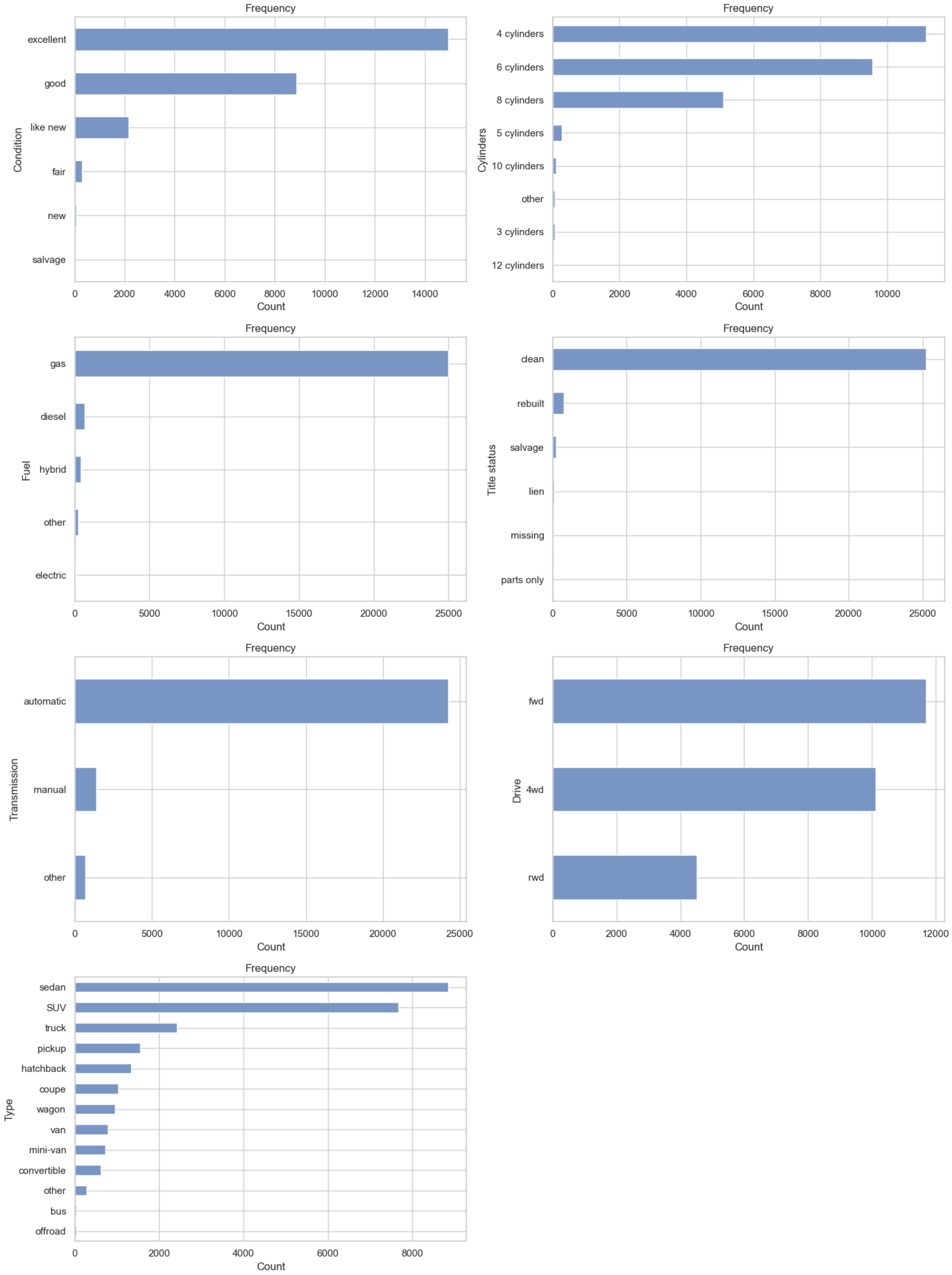


FIGURE 3: Frequency Diagrams of Dataset's categorical variables.

these categorical variables are grouped together to reduce data sparsity whilst improving the model's predictive performance by capturing these shared characteristics more effectively.

The categorical variables 'model' and 'manufacturer' are not presented in Figure 3, despite having numerous sparse categories. These variables are related in a way that diminishing the sparsity in 'model' can potentially reduce the sparsity in 'manufacturer'. Empirical evaluation of the model revealed that imposing a minimum threshold of 13 instances for the categories of the 'model' variable significantly enhances model performance with respect to the mean absolute error (MAE), hence this threshold was adopted.

Lastly, in the 'type' category we opted to removed vehicles classified as 'offroad' and 'bus' from our dataset because the characteristics of these specialized vehicles significantly set them apart from the rest of the entries in our dataset. Additionally, the vehicles classified under 'other' might exhibit a broad spectrum of characteristics. This variability could compromise the accuracy of the model, leading us to decide to exclude them as well.

	Price($)	Year	Odometer
Mean	11486	2011	119508
Std	5988	5	59976
Min	2350	1990	0
25%	6900	2008	84527
50%	9995	2012	116164
75%	14999	2014	150157
Max	27000	2020	2319010

TABLE II: Descriptive statistics for Price, Year, and Odometer.

Table II presents the descriptive statistics of our numerical variables after implementing the data cleaning procedures outlined earlier. Notably, even though we have excluded all listings classified as new, our dataset still includes some listings with an odometer value of 0. Such entries may represent listings with missing or erroneous information. Consequently, these have been removed from our dataset.

4.3 Visual Analysis of Price Trends and Geographic Distribution in the Used Car Market

Figure 4 is a line chart that visualizes the average prices of used car listings over the period starting in 1990 and ending in 2020. The highest average of approximately $35K is obtained in the most recent examined year, 2020, while the lowest is approximately $6K and is obtained in 1998. Additionally, the general trend for the confidence interval is to widen for years more distant from the present as there are fewer listings available for those years.

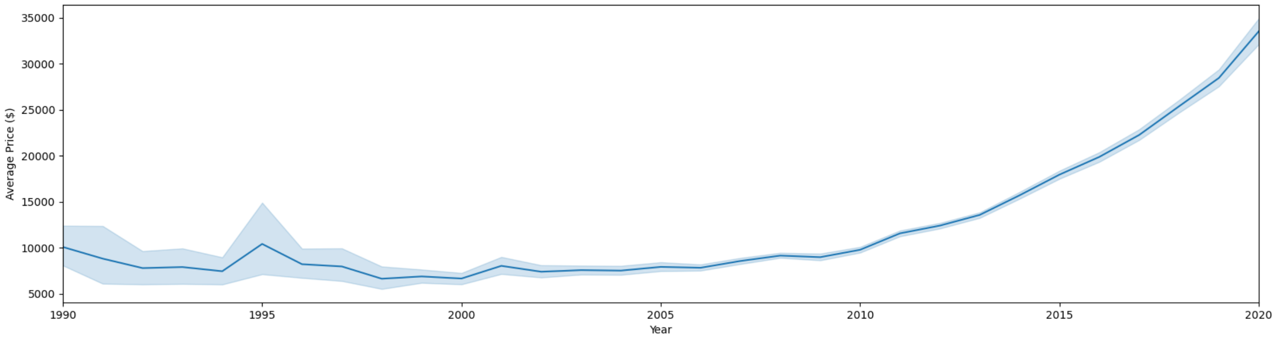


FIGURE 4: Horizontal Bar Plot Depicting the Year-wise Average Prices of Used Cars with Error Bars.

Figure 5 displays a visual representation of a random 5% sample of items from the dataset, plotted as red dots over the geographical layout of the United States. These dots provide a visual indication of the distribution and concentration of the sampled items across different regions of the United States.

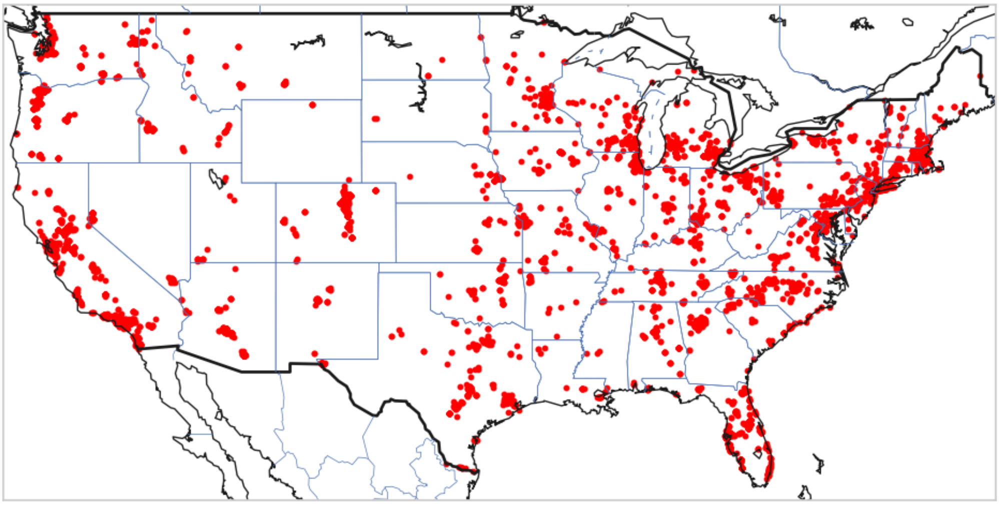


FIGURE 5: Geographic Distribution of a Random Sample.

The provided visual representation suggests that the samples in the dataset are geographically distributed with a substantial density variation. The highest concentration of red dots is visible in the eastern half of the country, particularly along the East Coast. The Western United States shows a sparser distribution, with red dots spread out more evenly, except for the populous states of California, Oregon, and Washington. The visualization excludes Alaska and Hawaii due to their limited data points, highlighting the focus of the dataset primarily on continental United States. Overall, there is a notable clustering of data points around major urban areas, which could imply that the dataset has a significant urban bias.

Figure 6 represents an extract of a heatmap. The heatmap visualizes the average price of the most popular car model in our dataset, the Ford F-150, across different geographic locations. Each location's intensity (or color gradient) is determined by the average price of the vehicle at that specific geographic coordinate. Therefore, regions on the map with warmer colors would suggest higher average prices for the F-150 model, whereas cooler colors indicate lower average prices.

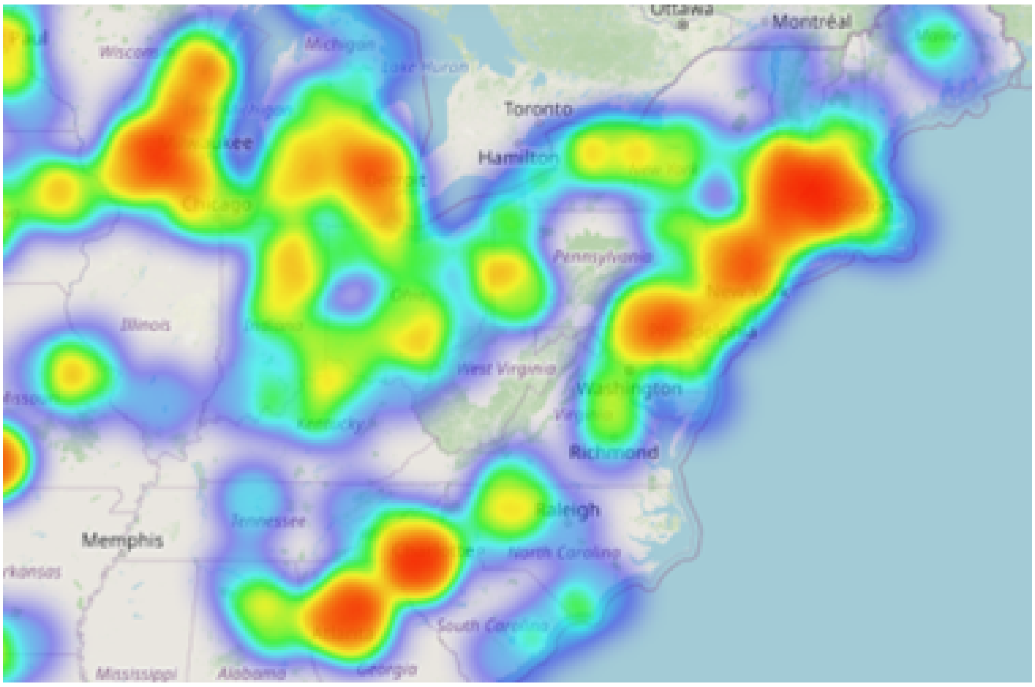


FIGURE 6: Extract of Heatmap Illustrating the Geographic Variation in Average Prices of the Ford F-150.

From the visualization, we can infer that the average price for the Ford F-150 is not uniform across the region. There are areas of high average prices along the Northeast Corridor. Conversely, in the complete heatmap, it can be observed that rural areas show cooler colors. In conclusion, pricing trends seem to diminish as one moves away from urban centers, which could suggest that urban markets have a positive influence on car pricing compared to rural areas.

4.4 Model Construction and Hyperparameter Optimization for Predictive Accuracy

The primary aim of this study is to construct a predictive model capable of capturing 90% of actual used car prices within its predicted intervals. We will evaluate the success of our model in meeting this objective using empirical methods.

To begin with, the dataset is randomly divided into two distinct sets: a training set S_1 and a test set S_2. The training set, which comprises 80% of the observations, is used for building the model while the test set, accounting for the remaining 20%, is utilized for fine-tuning and validating the accuracy of the model. For conformal models, out of the training set S_1, an additional 20% of the data points are randomly extracted to be held as calibration data, leading to them being trained with only 60% of the initial dataset. It is reasonable to suggest that non-conformal models, trained with more data, might produce more precise point predictions, particularly in the context of small datasets. However, the primary objective of this study is to derive prediction intervals that are reliably accurate, even if this means reducing the accuracy of point predictions.

To address the challenge of finite-sample variability, a rigorous approach was adopted. This involved running the model 100 times, each time using distinct sets for training, calibration, and testing. By doing so, the inherent variation due to the limited sample size can be mitigated. This methodological rigor enhances the robustness of the model against the fluctuations that typically arise from small sample sizes, thereby providing a more reliable and consistent analysis.

The model is based on a gradient boosting tree algorithm which is employed to fit the training data on two quantiles. Our quantile estimates are impacted by the hyperparameters of our algorithm. Thus, a grid search is first conducted to identify the optimal hyperparameters when minimizing the MAE on the validation data.

- Beyond a certain number of trees, there is a diminishing marginal effect on out-of-sample accuracy, meaning that adding more trees contributes less to the predictive capabilities of the model, and may even degrade performance due to overfitting (Palenicek et al., 2023). Therefore, the number of boosted trees to fit is limited to a set of specific values: $\{4000, 4500, 5000, 5500, 6000\}$. Eventually, 5000 was identified as the optimal value.
- To control model complexity, we impose limits on both the maximum number of tree leaves and the depth of the base learner trees. Specifically, the number of leaves per tree is constrained to one of the following values: $\{32, 64, 128, 256\}$. Similarly, the depth of each tree is restricted to $\{8, 16, 32, 64\}$. These restrictions are essential for maintaining a balance between model complexity and its ability to generalize. In our case, the optimal hyperparameters were, respectively, 64 and 8.
- The boosting learning rate, which influences the convergence speed and step size in minimizing residuals, is set to one of these values $\{0.01, 0.05, 0.1\}$. A carefully chosen boosting learning rate acts as a regularization term and helps to prevent the model from overfitting. In our scenario, the ideal boosting learning rate was 0.05.

The optimal hyperparameters identified as a result of the grid search where used to train our model.

4.5 Empirical Evaluation of Confidence Interval Prediction

We will now conduct an empirical evaluation of the FQR and CQR methods to determine their accuracy in predicting confidence intervals. Specifically, we aim to assess whether 90% of actual prices fall within the forecasted ranges provided by these methods.

In this process, we will evaluate the marginal coverage, defined as the coverage of prediction intervals of actual observations across the entire dataset. Marginal coverage does not consider how accuracy is distributed among different subgroups of our dataset (Angelopoulos & Bates, 2023). If the marginal coverage is greater than or equal to our desired nominal level, then the condition specified in Equation 1 is satisfied.

$$\text{Marginal Coverage} = \frac{1}{|I_{\text{test}}|} \sum_{i \in I_{\text{test}}} \mathbf{1}(Y_i \in \mathcal{C}(\boldsymbol{X}_i)) \tag{9}$$

In Equation 9, I_{test} is the set of indices associated with the test set and $|I_{\text{test}}|$ represents the number of indices. The summation $\sum_{i \in I_{\text{test}}}$ implies that we are considering each used car i in the test dataset. Furthermore, this research examines the breadth of prediction intervals. We measure the width of these intervals relative to the median rather than the mean to diminish the impact of the skewness of our dataset.

$$\text{Median Width} = \text{median}\left(\frac{\left|q_{1-\frac{\alpha}{2}}(X_{n+1}) - q_{\frac{\alpha}{2}}(X_{n+1})\right|}{Y_{n+1}}\right) \tag{10}$$

Table III demonstrates a trade-off between accuracy and precision in price prediction methods. While CQR method provides broader prediction intervals resulting in a marginal coverage equal to the desired nominal level, the FQR method offers narrower but less inclusive predictions.

	Marginal coverage	Median width
CQR	0.90	0.70
FQR	0.72	0.45

TABLE III: Marginal Coverage and Median Width at 90% Nominal Coverage Probability.

The marginal coverage of conformal prediction intervals for the continuous features of our model (‘odometer’, ‘year’, ‘latitude’, and ‘longitude’) is visualized in the charts below:

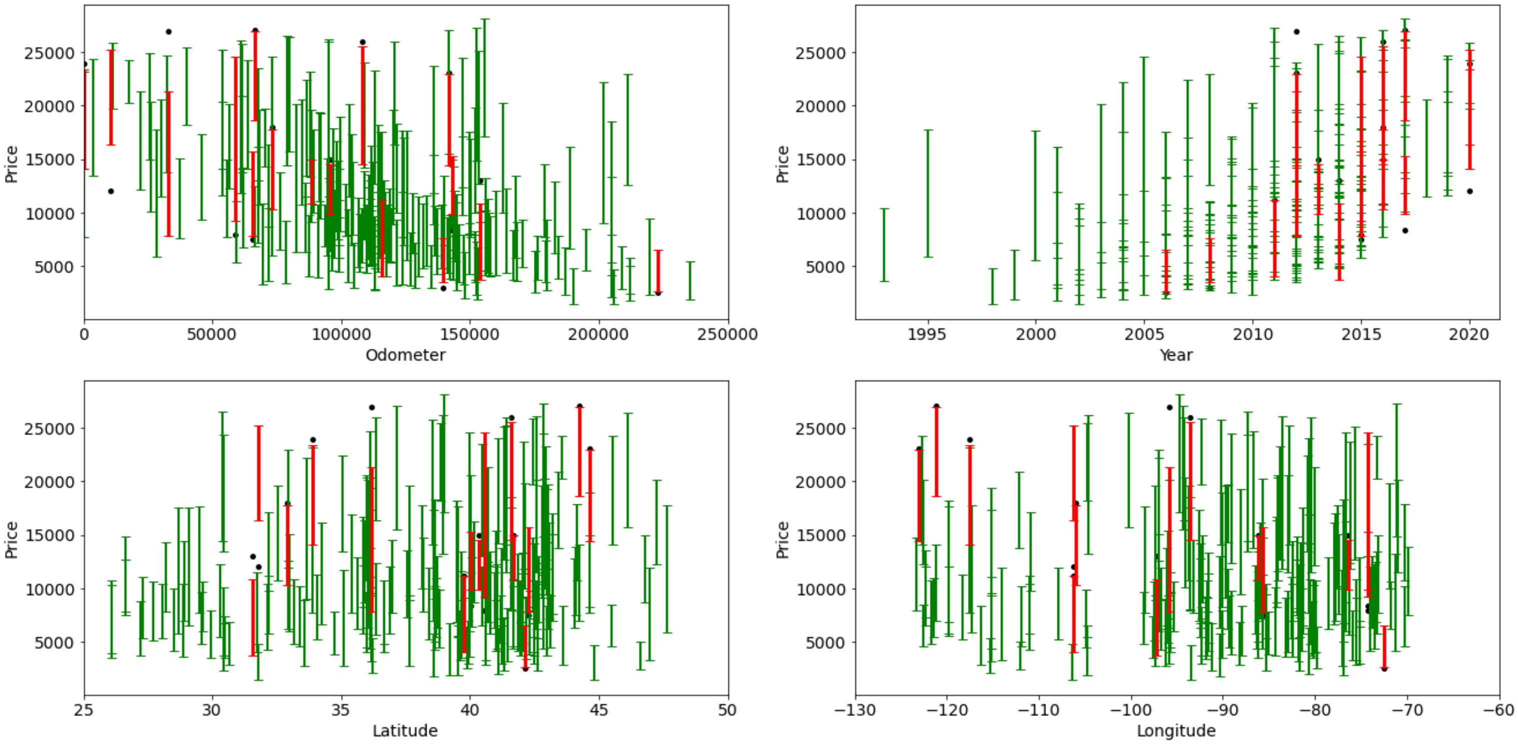


FIGURE 7: Visualization of Prediction Intervals for Selected Features at 90% Nominal Coverage Probability.

In these charts, green bars represent cases where the listing prices fall within the predicted intervals, while red bars denote instances where they lie outside these intervals. Additionally, when the real price is not within the interval, a black dot is included to mark the listing value of the used car.

Since conformal predictions are bound by Equation 1, the proportion of prediction intervals that fail to include the actual prices must be smaller than or equal to the pre-established significance level, denoted as α, of 10%. On the other hand, the proportion of prediction intervals that do encompass the true prices must be bigger than or equal to the NCP of 90% [1].

4.6 Analysis of Conformal Quantile Regression Performance Across Subgroups with Conditional Coverage

Building on the marginal coverage concept discussed in Section 2, coverage has been evaluated across the entire dataset up until now. By treating all variables as random, including $(\boldsymbol{X}_{n+1}, Y_{n+1})$ and the data for training our machine learning model, we lack assurance that the interval will cover (Y_{n+1}) when conditioned on a specific observed value of $\boldsymbol{X}_{n+1}$, as noted by Sesia & Candès (2020). Conditional coverage provides a more focused assessment, examining how well the model performs across various subgroups in the dataset. This approach makes sure that the accuracy of the model is not only high overall, but also homogeneous across different segments. For instance, in our use case, measuring conditional coverage can expose that our model excels with gas cars but underperforms with diesel cars. In mathematical terms, the conditional coverage guarantee is:

$$\Pr(Y_{n+1} \in \mathcal{C}(\boldsymbol{X}_{n+1}) | \boldsymbol{X}_{n+1} = \boldsymbol{x}) \geq 1 - \alpha \tag{11}$$

that should hold for any $\boldsymbol{x}$.

To obtain valid conditional coverage across all of our subsets, it would be necessary to train separate models for each specific region, for example gas and diesel cars, and then conformalize the intervals. Given that this is not implemented in our model, achieving valid conditional coverage would require us to make strong assumptions about the joint distribution $P_{X,Y}$ that, as Barber et al. (2021) points out, rarely hold.

Therefore, although the conditional coverage guarantee does not apply, conditional coverage values are still potentially informative. To calculate conditional coverage, the dataset was first divided in tertiles as shown in Table IV.

Variable	1st Tertile	2nd Tertile	3rd Tertile
Odometer	< 9600	$9600 \leq \textit{ODO} < 136555$	≥ 136555
Year	< 2009	$2009 \leq \textit{YR} < 2013$	≥ 2013

TABLE IV: Bin cuts for Conditional Coverage at 90% Nominal Coverage Probability.

Dividing our dataset in tertiles resulted in the conditional coverages and median widths reported in Table V.

[1] For all graphs contained in Figure 7, out of the 200 depicted prediction intervals 16 don't contain the real listing price.

	Conditional coverage			Median width		
Variable	1st Tertile	2nd Tertile	3rd Tertile	1st Tertile	2nd Tertile	3rd Tertile
Odometer	0.76	0.90	0.91	0.78	0.70	0.83
Year	0.91	0.91	0.88	0.93	0.71	0.57

TABLE V: Conditional Coverage and Marginal Widths for Tertiles at 90% Nominal Coverage Probability.

Upon examining the conditional coverage for the 'odometer' variable in Table IV, we immediately notice that the value for the first tertile significantly deviates from the anticipated nominal coverage. This deviation confirms the lack of conditional coverage guarantees by conformal prediction. Further analysis of the remaining conditional coverage metrics reveal a general tendency for these values to oscillate around the NCP. In terms of median widths, it is observed that on average they are broader than those resulting from marginal coverage calculations, with the peak reached for 'year' variable in the 1st Tertile. This is expected given the modest number of samples available annually in this subset, as illustrated in Figure 4.

4.7 Extending the Empirical Analysis: Comparing Quantile Regression Methods at Varied Nominal Coverage Probabilities

In the previous sections, we focused on evaluating FQR and CQR methods, specifically targeting a 90% NCP. This evaluation revealed that while CQR successfully achieves the set coverage level, it tends to create prediction intervals that are wider than those of FQR. Moving forward, we plan to extend our empirical analysis to investigate the performance of these methods under a lower and higher NCP. The aim of this expanded study is to gain deeper insights into the balance between coverage level and interval width. By varying the nominal coverage, we will be able to better understand how adjustments in the NCP affect the trade-off between the width of prediction intervals and the marginal coverage of the intervals.

	80% Nominal Level		95% Nominal Level	
	Marginal Coverage	Median Width	Marginal Coverage	Median Width
CQR	0.80	0.50	0.95	0.92
FQR	0.59	0.31	0.81	0.59

TABLE VI: Marginal Coverage and Median Widths at 80% and 95% Nominal Coverage Probabilities.

As expected, at both 80% and 95% NCP, the CQR method demonstrates a marginal coverage matching the targeted level, with a median width significantly reduced at the 80% compared to both the 90% and 95% NCP scenario [2].

[2] Visualizations of the Conformal Prediction Intervals for the selected features at 80% and 95% NCP can be found in Appendix A and B.

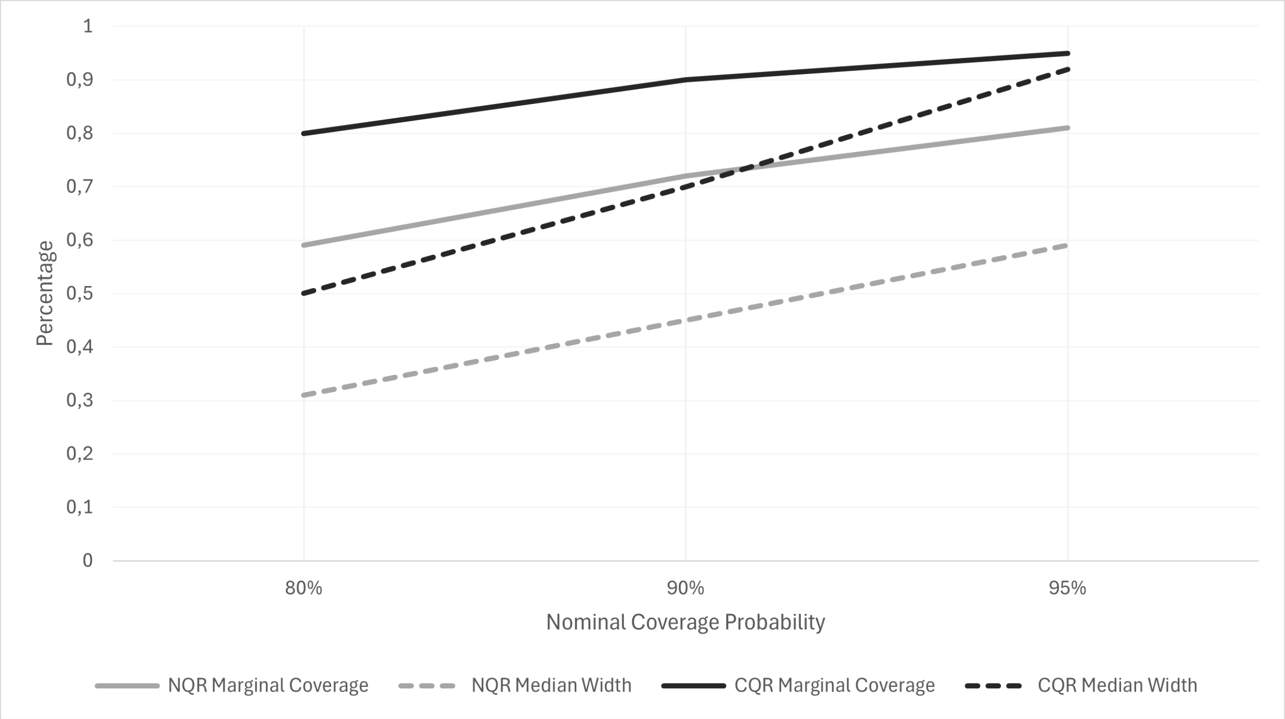


FIGURE 8: Marginal Coverage & Median Width for Normal Quantile Regression and Conformal Quantile Regression over Nominal Coverage Probability.

Figure 8 is a line plot that compares the performance of FQR and CQR across the different levels of NCP evaluated in this paper. The figure illustrates that the growth in median width is notably larger for CQR. Additionally, the marginal coverage of the FQR method approaches that of CQR as the NCP increases. These findings suggest that as the NCP level rises, FQR becomes increasingly more convenient when compared to CQR in both median width and marginal coverage, despite not reaching the NCP.

These insights are crucial when choosing among quantile regression methods for forecasting used car price ranges, with the choice significantly affected by the target NCP level. For instance, at an 80% NCP level, CQR is more advantageous compared to FQR than at a 95% NCP level. This is due to the fact that with rising NCP levels, CQR requires a considerable increase in the median interval width to achieve a less than proportional improvement in marginal coverage.

5 CONCLUSIONS

This study has successfully demonstrated the application of CP in a model for predicting used car prices. The integration of CP into a ML model has provided a significant advancement in the used car pricing domain. It has been particularly effective in generation predication intervals that are guaranteed to contain the actual listing value at the desired nominal value, thereby realistically representing the uncertainty in price estimations. This was possible thanks to the unique characteristic of CP needing not to rely on any assumption regarding the distribution of the dataset, which is particularly beneficial in the used car market where variability and market fluctuations are common.

Our LightGBM model was first optimized for our dataset by minimizing the MAE and then the accuracy of prediction intervals was empirically assessed both for Normal and CQR. This evaluation showed a balance between interval coverage and width. While CQR uniquely maintained coverage at the desired nominal level, it resulted in a broader median width compared to FQR. This indicates that while CQR is more likely to cover the true parameter values, it does so with less precision compared to FQR. When implementing these techniques, users must weigh the importance of coverage against precision. They need to determine whether it is more crucial for their specific application to ensure a higher likelihood of encompassing the true value or to obtain a narrower interval estimate.

Subsequently, conditional coverage was evaluated empirically through marginal coverage. The study reveals that CQR can fall short of achieving the expected conditional coverage levels. We observed a particularly pronounced discrepancy in the 1^{st} tertile of the odometer variable, where we found a gap of 14% percentage points between the actual conditional coverage and the targeted nominal level.

Following this, the comparison of quantile regression methods at varied nominal coverage probabilities revealed that to maintain NCP level, CQR tends to increase the median interval width more than proportionally. Conversely, CQR, despite not reaching the target NCP, becomes relatively more efficient in terms of median width and marginal coverage at higher NCPs. These insights are crucial for selecting appropriate quantile regression methods for forecasting, highlighting the importance of considering the target NCP level as a factor in the decision-making process.

Moreover, the study uncovered that the median width of the prediction intervals increases significantly when the number of training observations is smaller. This insight opens avenues for future research, where a dataset without the limitations highlighted in Section 4.3 could be studied, thereby enhancing the accuracy of prediction intervals across the whole dataset.

References

Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, *16(4)*, 494-591.

Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, *10(2)*, 455-482.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, *63*(4), 841–890.

Breiman, L. (2001). Random forests. *Machine Learning*, *45(1)*, 5-32.

Chen, C., Hao, L., & Xu, C. (2017). Comparative analysis of used car price evaluation models. *AIP Conference Proceedings*, *1839(1): 020165.*

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189-1232.

Grand View Research. (2022). *Used car market size, share, growth & trends report, 2030.*

Griliches, Z. (1961). Hedonic price indexes for automobiles: An econometric analysis of quality change. *The Price Statistics of the Federal Government*, *73*, 173–196.

Han, S., Qu, J., Song, J., & Liu, Z. (2022). Second-hand car price prediction based on a mixed-weighted regression model. *2022 7th International Conference on Big Data Analytics (ICBDA)*, 90-95.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46(1)*, 33-50.

Lei, J., & Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *76*, 71-96.

Ozgur, C., Hughes, Z., Rogers, G., & Parveen, S. (2016). Multiple linear regression applications automobile pricing. *Internatinal Journal of Mathematics and Statistical Invention(IJMSI)*, *4(5)*, 13-20.

Palenicek, D., Lutter, M., Carvalho, J., & Peters, J. (2023). Diminishing return of value expansion methods in model-based reinforcement learning. *The Eleventh International Conference on Learning Representations.*

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. *In Proceedings of Machine Learning: European Conference of Machine Learning*, 345-356.

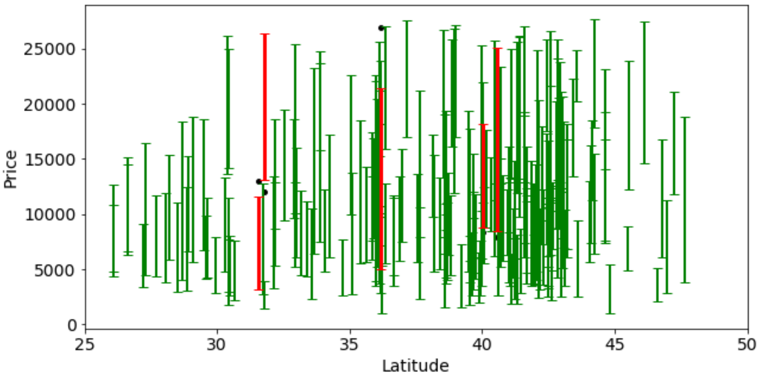
Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems*, *32*, 3543-3553.

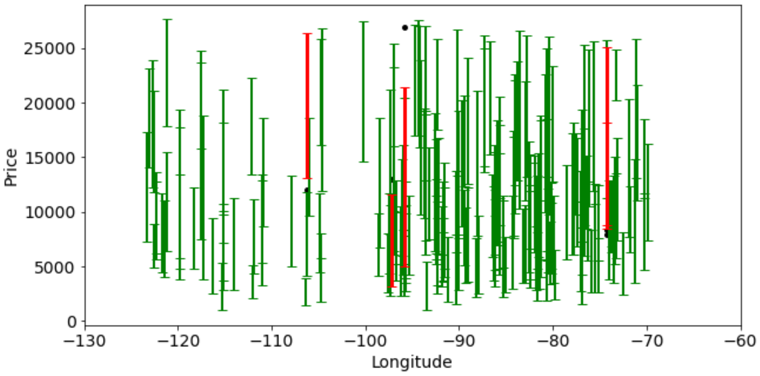
Sesia, M., & Candès, E. J. (2020). A comparison of some conformal quantile regression methods. *Stat*, *9(1)*, e261.

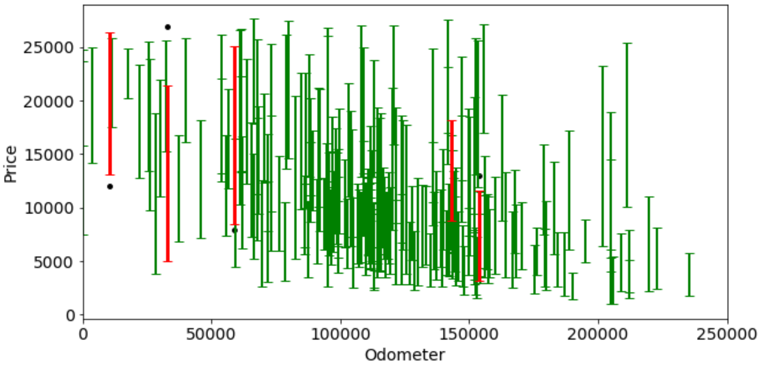
Varshitha, J., Jahnavi, K., & Lakshmi, C. (2022). Prediction of used car prices using artificial neural networks and machine learning. *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 1-4.

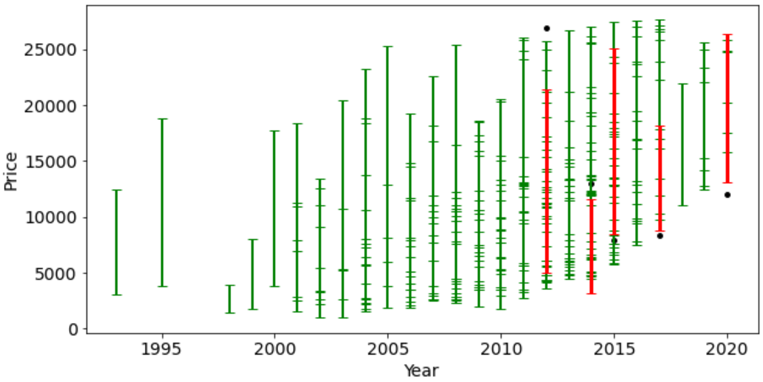
Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world.*

A Visualization of Conformal Prediction Intervals for Selected Features at 95% Nominal Coverage Probability.

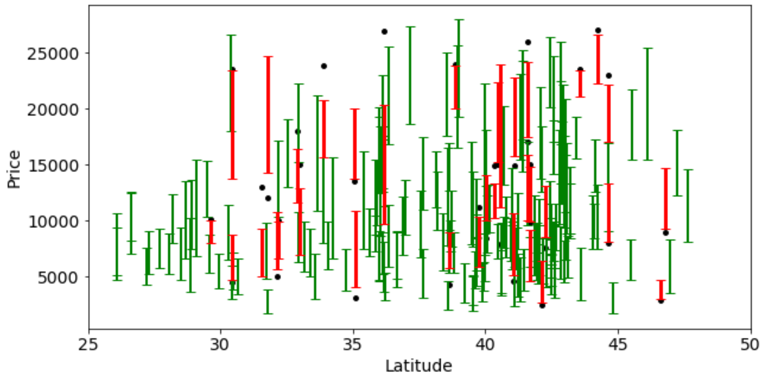


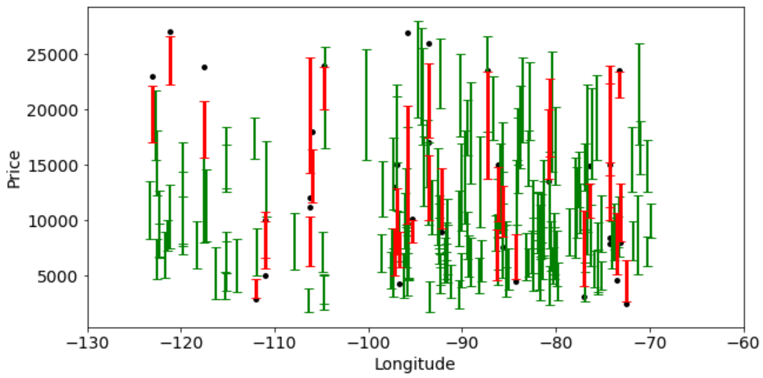


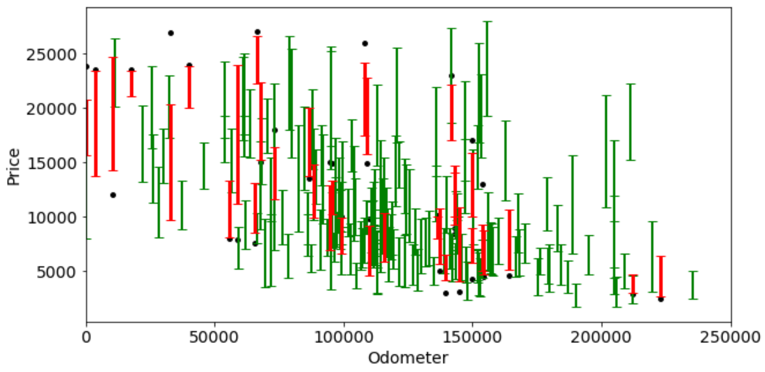


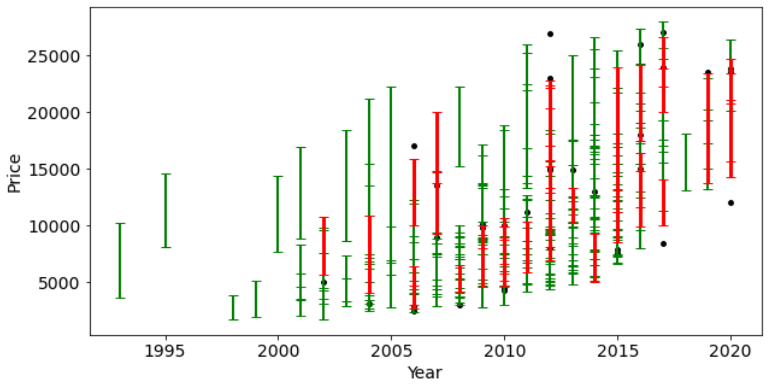


B Visualization of Conformal Prediction Intervals for Selected Features at 80% Nominal Coverage Probability.


Price
Latitude


Price
Longitude


Price
Odometer


Price
Year