MASTER
APPLIED ECONOMETRICS AND FORECASTING

MASTER´S FINAL WORK
DISSERTATION

CONFORMAL PREDICTION OF REAL ESTATE PRICES WITH MACHINE LEARNING

JEANNE PAQUETTE

SUPERVISION:
JOÃO AFONSO BASTOS

OCTOBER - 2023

To my family and friends for
your constant support
throughout my academic
journey and specifically
during the writing of this
dissertation.

GLOSSARY

PCA – Principal Component Analysis
AVM – Automated valuation model
ANN – Artificial Neural Network
MAE – Mean Absolute Error

RESUMO

A quantificação da incerteza associada à avaliação imobiliária tem sido notavelmente negligenciada na literatura. O objetivo deste trabalho é colmatar a lacuna existente, mediante uma análise da incerteza na avaliação de propriedades, através da aplicação de técnicas de aprendizado de máquina e previsão conformal. A previsão conformal quantifica a incerteza associada a previsões individuais e proporciona uma série de resultados possíveis em torno de estimativas pontuais com base em um nível de significância pré-definido. Ao aplicar a regressão de quantis conformal, somos capazes de mitigar as limitações das abordagens iniciais de regressão conformal e construir intervalos que exigem apenas que os dados sejam passíveis de intercâmbio para assegurar a cobertura. Através de um estudo empírico dos preços de imóveis na área da Baía de São Francisco, descobrimos que a regressão de quantis conformal fornece intervalos de previsão adaptativos com cobertura garantida que capturam variações inerentes à incerteza observada entre distintos níveis de preços de propriedades.

PALAVRAS-CHAVE: Avaliação de imóveis; Predição conforme; Aprendizado de máquina; Regressão quantílica conforme

ABSTRACT

Uncertainty quantification associated with real estate appraisal has largely been ignored in the literature. The aim of this dissertation is to fill this gap by analysing uncertainty in property valuation using machine learning complemented by conformal prediction. Conformal prediction quantifies uncertainty associated with individual predictions and provides a range of possible outcomes around point estimates based on a pre-defined significance level. By applying conformal quantile regression, we can mitigate limitations of early conformal regression approaches and we are able to build intervals that only require the data to be exchangeable for the coverage to be guaranteed. Through an empirical study of property prices in the San Francisco Bay Area, we find that the conformal quantile regression provides adaptive prediction intervals with guaranteed coverage that captures uncertainty variations across different property prices.

KEYWORDS: Property valuation; Conformal prediction; Machine learning; Conformal quantile regression

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I wish to thank my supervisor, Professor João Afonso Bastos, for presenting this very interesting research topic to me, and for his consistent guidance throughout this project.

I would like to thank my colleagues with whom I got to work on projects with, share ideas and study notes, and have interesting discussions regarding the study material.

I am also forever grateful to my family and friends for their patience, relentless support, and for always believing in my abilities to complete this master’s degree.

1. INTRODUCTION

Real estate transactions are a key contributor to the overall economy, creating employment, generating tax revenues, and stimulating economic growth. A strong real estate market, where both demand and property values are high and constant, is a great indication of a robust economy and positive consumer confidence. This, in turn, affects consumer spending, investment levels and overall economic certainty. The housing market also plays an important role in the livelihood of economic agents, namely as a primary source of wealth creation. Real estate investment represents the most substantial investment most individuals undertake in their lifetime. Consequently, housing market conditions, as well as housing affordability and availability, greatly impacts the quality of life and standards of living for most. Moreover, the real estate market is also used for many as a hedge against inflation, either by direct purchase or by including this asset class into a well-diversified investment portfolio. To offer effective guidance and advice to prospective home buyers, investors, and property sellers alike, real estate brokers and investment advisers rely on predictive tools such as automated valuation models (AVMs). This valuation method allows them to estimate property values, understand the effect of specific property attributes on pricing, and provide well-informed recommendations to their clients (Bellotti, 2017).

The conventional method for estimating house prices is hedonic in nature, i.e., based on the dwellings' attributes and follows the widely used linear regression technique (Goodman, 1978). For the model to provide valid coefficient estimates, i.e., unbiased, consistent, and efficient, underlying assumptions must be satisfied such as linearity, exogeneity of the variables, and independence of the errors. More precisely, the violation linearity and exogeneity can produce both bias and inconsistent coefficient estimators. Although problems like autocorrelation of errors and heteroskedasticity—identified when the variance of the error term is not constant—do not result in biased estimators, they do yield inefficient parameter estimates and compromise the dependability of hypothesis tests. Considering the nature of housing data, those assumptions may be easily violated which reduces the precision of linear models and can lead to invalid inference (Limsombunchai, et al., 2004).

In recent years, with the increasing access to large databases, the use of machine learning models—decision trees, random forest, neural networks, and gradient boosting—are becoming commonly used to perform this task. In contrast to linear models, machine learning algorithms are less dependent on satisfying the previously stated assumptions. Artificial neural networks

(White, 1989), support vector machines (Kecman, 2005) and tree-based models such as random forest (Breiman, 2001) and gradient boosting (Freund & Schapire , 1996; Friedman, 2001) are common predictive methods, gradually replacing the linear hedonic models. These methods are increasingly employed as they can handle complex, unstructured, and high-dimensional data. Moreover, they have a strong ability to recognize and model nonlinear patterns making them more suitable to perform property valuations (Babu & Chandran, 2019).

The primary objective of this dissertation is to contribute to the growing area of research focused on real estate market valuation using machine learning algorithms and address the current gap in the literature regarding uncertainty quantification. To achieve this, the dissertation employs conformal quantile regression in conjunction with a machine learning model to generate prediction intervals. These intervals serve as a tool to quantify the uncertainty associated with property price predictions. Specifically, we investigate the performance of a conformal quantile model in comparison to a traditional approach of standard quantile regression. The study focuses on data from the dynamic and highly competitive real estate market of the San Francisco Bay Area, where property valuations are often characterized by complex and non-standard patterns. The dataset used for the empirical analysis contains 8,351 observations on sold properties, including their price, as well as their physical, location and geographical attributes. The houses comprised in the dataset were sold over the period 2020 to 2023. Considering that large datasets on sold properties are privately owned by real estate companies, the data used was acquired through an Application Programming Interface (API) service, and further enhanced using the United States Census Bureau website for geographical and demographic data. The machine learning algorithm used to perform the analysis is the gradient boosting tree for quantiles. The research illustrates the advantages of using conformal prediction as it outperforms conventional quantile regression, particularly in cases where the data displays strong variability in the prices. Notably, we find that conformal quantile regression provides adaptive prediction intervals with guaranteed marginal coverage centered at the desired nominal level of 0.90. The relative median widths reported also illustrates the sensitivity of the model to price variations.

The remainder of this dissertation is structured as follows: Section 2 covers the literature review, while Section 3 defines the conformal prediction model in the context of property price estimation. Section 4 presents the data used in the empirical study and provides an evaluation of the results. Concluding the dissertation, Section 5 provides a summary of the results from this analysis.

2. Literature Review

2.1 Hedonic price model

Most empirical research within the field of real estate economics tackle the analysis of housing price determinants. The primary objective is to identify the most effective model specifications that provide precise property valuation and predictive insights into trends in the real estate market. The foundational framework for the majority of empirical research, typically relies on the hedonic price model first introduced by Court (1939) and later popularized by Griliches (1971) and Rosen (1974). The hedonic model, applied to tangible assets, suggests that the price of a good is determined by the utility derived from its attributes and compares the price of related products with their individual characteristics. Thus, the price differentiation provides insights regarding the utility that consumers derive from the attributes and in turn justifies the price assigned to the good (Rosen, 1974).

Rosen (1974) introduces the pricing function:

$$p(\boldsymbol{x}) = p(x_1, x_2, \dots, x_n),$$

where $\boldsymbol{x}$ is the vector of characteristics which determines the price of the product. This simplistic hedonic model is widely used for real estate appraisal and market trend research. By applying this model to real estate valuation, we can estimate the value of a property by breaking down its various attributes such as the number of bedrooms, number of bathrooms, total square footage, location factors, etc., and the estimated price is thus determined considering the utility consumers derive from those (Chau & Chin, 2003).

Following the hedonic framework, model specification for real estate valuation includes key housing characteristics, including the number of bedrooms and bathrooms, total square footage of the house and lot, availability of parking space, the presence of a fireplace, and more. As evidenced by Can (1992), these attributes exhibit a positive correlation with housing prices. Hence, hedonic theory suggests their inclusion for property valuation as they constitute the set of implicit prices affecting the overall value of a home (Limsombunchai, et al., 2004). Can's study not only underscores the relevance of structural attributes but also emphasizes the crucial role of socio-economic factors in assessing the impact of neighbourhood attributes on property valuation. By incorporating variables such as the percentage of nonwhite residents, poverty rate, unemployment rate, median household income, and the percentage of vacant units, Can constructs a neighbourhood quality index using Principal Component

Analysis (PCA). As anticipated, the study reveals a positive correlation between housing prices and the neighbourhood quality index, indicating that an improvement in the quality of the area evaluated is associated with an expected increase in property values.

Considering these findings, our analysis integrates socio-economic variables to enhance the dataset's quality, ultimately improving the predictive accuracy of our model. This holistic approach considers both structural housing attributes and socio-economic factors, providing a comprehensive foundation for real estate valuation in our study.

2.2 Machine learning models

In terms of model selection, most of the recent literature analyzing the prediction of house prices establishes that machine learning algorithms outperform the traditional linear hedonic regression model used for real estate evaluation (Hjort, et al., 2022; Wang & Wu, 2018; Peter, et al., 2020). A study by Grudnitski and Do (1992) demonstrates that predictions made using artificial neural networks (ANN) yield more predictive accuracy than those generated by linear regression models. This finding highlights the suitability of ANN for real estate valuation considering their high ability to handle complex and noisy data. A recent study by Ho et al. (2021) which compares three machine learning algorithms, namely support vector machines, random forest, and gradient boosting machines, suggests that both tree-based models – random forest and gradient boosting machine – outperformed by a significant margin the support vector machine algorithm.

Although the literature suggests that neural networks and tree-based models provide more prediction accuracy over the linear model (McCluskey, et al., 2012; Hjort, et al., 2022), it neglects an important shortcoming of machine learning prediction. Uncertainty quantification is precisely what is missing from the literature in the context of real estate appraisal using machine learning techniques. The use of uncertainty quantification has far-reaching implications for real estate professionals, investors, and stakeholders alike, considering it facilitates risk-prediction for lending decisions and enables investors to make well-informed decisions given certain market conditions. The relative difficulty or ease of pricing for certain properties may be better understood using prediction intervals, with their widths being a barometer of pricing confidence. Narrower intervals suggest more precise property valuations and larger ones indicate pricing difficulty making the prediction less reliable and should incite the client to request a revaluation. Moreover, mortgage credit risk can be moderated by using the lower bound of the prediction interval as a conservative estimate for predicted property prices (Bellotti, 2017).

2.3 Approaches to uncertainty quantification

To build the prediction interval $C(\boldsymbol{X}_{n+1}) \subseteq \mathbb{R}$ for new properties on the market without a valuation price Y_{n+1}, we must consider a vector of n property attributes $\boldsymbol{X}_{n+1}$, such as number of bedrooms, number of bathrooms, year built, and lot size, and $\{Y_i\}_{i=1}^{n}$ known property prices for each $i \in \{1, \ldots, n\}$. The aim is thus to obtain a prediction interval that has a high probability of containing the unknown valuation price Y_{n+1} conditional on $\boldsymbol{X}_{n+1}$.

$$\mathbb{P}\{(Y_{n+1} \in C(\boldsymbol{X}_{n+1}) | \boldsymbol{X}_{n+1}\} \geq 1 - \alpha. \qquad (1)$$

where α is the pre-defined error rate, and $1 - \alpha$ represents the nominal coverage rate.

One way of constructing prediction intervals involves using an estimate of the standard deviation $\hat{\sigma}(x)$ as the uncertainty measure. Assuming $Y \mid \boldsymbol{X} = x$ follows a Gaussian distribution, we can generate the mean and variance of a trained model which follows the same parametric distribution as x. By maximizing the likelihood function with respect to both the mean and variance of the data generating process, we obtain the point estimate and its associated standard deviation, which is used as the measure of dispersion and thereby uncertainty. That is, $\hat{\sigma}(x)$ will be small if the model properly captures the price variations and large otherwise. Although this approach is widely used in statistics and machine learning to quantify uncertainty, it is not reliable considering the distributional assumptions are not always satisfied (Angelopoulos & Bates, 2023).

In machine learning, the ensemble method technique to measure uncertainty is also widely used. The goal is to generate an ensemble of models, each varying slightly, and the spread of the property price predictions is then used to quantify uncertainty. Techniques such as Bagging, Boosting, and Random Forest are common ensemble methods used to train the models, and each represent a different approach to predicting property prices (Ho, et al., 2021). However, ensemble methods require re-training and maintaining multiple models which increases the memory requirements and computational complexity of using this technique, especially with large ensembles.

Hu et al. (2022) proposed to mitigate the increased complexity issue by using dropout neural networks which involves randomly dropping out a portion of the nodes during the training as well as the prediction phases. By doing so, we can capture the model's uncertainty by analyzing the dispersion between the ensemble made of different network configurations. As the authors point out, this method also has drawbacks considering it depends heavily on the choice of the dropout rate for each forward and backward step. In cases where the dropout rate is too low or the network is large, the estimation tends to underestimate the associated uncertainty measure. As well, the dropout rate is not inherently calibrated with the confidence level, indicating that the resulting prediction intervals are not necessarily valid (Hu, et al., 2022).

A solution to unreliable prediction intervals is the use of quantile regression (Koenker & Bassett, 1978), which allows us to estimate conditional prediction intervals of property prices Y_{n+1} that are robust to heteroskedasticity (Feldman, et al., 2021). The conditional quantile function is defined by Romano et al. (2019) as

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

where $F(Y|\boldsymbol{X})$ denotes the conditional distribution function of property prices Y given the vector of attributes $\boldsymbol{X}$. Let $q_{\alpha_L}(\boldsymbol{X})$ and $q_{\alpha_H}(\boldsymbol{X})$ denote the lower and upper conditional quantiles, respectively, where $\alpha_L = \alpha/2$ and $\alpha_H = 1 - \alpha/2$. The conditional quantile prediction interval for Y_{n+1} is thus given by:

$$C(\boldsymbol{X}) = \left[q_{\alpha_L}(\boldsymbol{X}_{n+1}), q_{\alpha_H}(\boldsymbol{X}_{n+1})\ \right] \tag{3}$$

and theoretically satisfies Equation (1).

By minimizing the pinball loss function denoted as:

$$\rho_L(Y, \hat{Y}) = \begin{cases} \alpha(Y - \hat{Y}) & if\ Y - \hat{Y} > 0, \\ (1 - \alpha)(Y - \hat{Y}) & otherwise. \end{cases} \tag{4}$$

we can estimate $q_{\alpha}(\boldsymbol{X})$ that best captures the conditional distribution of property prices based on the housing characteristics.

An issue that arises while employing quantile regression to obtain valid coverage comes from the fact that this method does not provide finite-sample guarantee (Romano, et al., 2019). Hence, given the strong distributional assumptions that must be satisfied for the conditional coverage to hold, in practice, the quantile interval in Equation (3) rarely guarantees the desired coverage in Equation (1), and frequently leads to an under-coverage bias (Bai, et al., 2021). A way to mitigate this issue involves the use of *conformal prediction*, which allows us to relax distributional assumptions and build prediction intervals with finite-sample marginal coverage guarantee given by:

$$\mathbb{P}\{(Y_{n+1} \in C(\boldsymbol{X}_{n+1})\} \geq 1 - \alpha. \tag{5}$$

This is achieved assuming only exchangeability of the data, i.e., drawn i.i.d. from a joint distribution $P_{\boldsymbol{X},Y}$ (Romano, et al., 2019; Angelopoulos & Bates, 2023). This model-agnostic method is useful in the context of real estate valuation considering it provides reliable prediction intervals, enhancing transparency in the appraisal process.

3. Conformal prediction of property prices

In the context of real estate appraisal, we consider the price of property as the dependent variable Y, and $\boldsymbol{X}$ as the vector of property attributes. To build our prediction interval $C(\boldsymbol{X}_{n+1}) \subseteq \mathbb{R}$ on new test data for properties with known attributes $\boldsymbol{X}_{n+1}$[1], but unknown valuation price Y_{n+1} we train our machine learning model using a sample of n $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$. The only assumption necessary to guarantee coverage is for all the samples $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n+1}$ to be exchangeable and drawn from a joint distribution $P_{\boldsymbol{X},Y}$. Following this framework and given any nominal coverage rate $1 - \alpha$, the constructed prediction interval $C(\boldsymbol{X}_{n+1})$ is said to satisfy Equation (5).

Following the split method by Papadopoulos et al. (2002) we build prediction intervals by splitting the training set into two subsets: the training set $S_t = \{(\boldsymbol{X}_i, Y_i) : i \in I_1\}$ and the calibration set $S_c = \{(\boldsymbol{X}_i, Y_i) : i \in I_2\}$. We first fit a regression model $Y = f(\boldsymbol{X})$ on the training set S_t:

$$\hat{f}(\boldsymbol{X_i}) \leftarrow A\,\{(\boldsymbol{X}_i, Y_i) : i \in I_1\}, \tag{6}$$

where A may be any regression algorithm since split conformal prediction does not require any distributional assumptions, even exchangeability of the data.

Then, to obtain the conformity scores we calculate the absolute residuals of the trained model on the calibration set:

$$\hat{\varepsilon}_i = \left|Y_i - \hat{f}(\boldsymbol{X_i})\right|, i \in I_2. \tag{7}$$

[1] Y_{n+1} refers to the test response and $\boldsymbol{X}_{n+1}$ refers to the test data.

Based on a pre-defined significance level α and n_2 observation in the calibration set, we can compute the quantile $q_{1-\alpha}\,(\hat{\varepsilon}_i, I_2)$ of the empirical distribution of the absolute residuals,

$$q_{1-\alpha}\,(\hat{\varepsilon}, I_2) = \frac{(n_2+1)(1-\alpha)}{n_2} \text{ empirical quantile of } \hat{\varepsilon}_i : , i \in I_2. \tag{8}$$

Finally, the prediction interval for the property price Y_{n+1} with attributes $\boldsymbol{X}_{n+1}$ is given by:

$$C(\boldsymbol{X}_{n+1}) = \left[\hat{f}(\boldsymbol{X}_{n+1}) - q_{1-\alpha}\,(\hat{\varepsilon}, I_2), \hat{f}(\boldsymbol{X}_{n+1}) + q_{1-\alpha}\,(\hat{\varepsilon}, I_2)\,\right]. \tag{9}$$

Although this prediction interval is guaranteed to satisfy Equation (5), a major limitation is found from the fact that the width of the interval is fixed and independent of $\boldsymbol{X}_{n+1}$ (Romano, et al., 2019). Hence, the prediction interval is not adaptive, in that the property prices variations observed are not captured using the split method.

Early versions of conformal prediction faced some limitations in terms of adaptivity to variation of the dependent variable. That is, early methods provided prediction intervals that have fixed interval width, which is restrictive in the context of property prices considering the significant discrepancies observed in the market. By applying *conformal quantile regression*, we can mitigate limitations of early conformal regression approaches and we can build intervals that are well-calibrated and adaptive to variation across different property prices and defined attributes.

3.1 Conformalized quantile prediction

The framework of conformal quantile prediction, detailed by Romano et al. (2019), is similar to the splitting method in that we must also divide the dataset into two subsets: the estimation set S_t, and the calibration set S_c used to calculate the conformity scores, which contains 80% and 20% of the observation, respectively. After setting the error rate α at 0.10, such that 90% of the actual property prices fall within the constructed conformal interval, we use any quantile regression algorithm A, to train on S_t the lower and upper conditional quantiles,

$$\left\{\,\hat{q}_{\alpha_L}(\boldsymbol{X}), \hat{q}_{\alpha_H}(\boldsymbol{X})\right\} \leftarrow A\,\{(\boldsymbol{X}_i, Y_i) :\ i\ \in\ I_1\}, \tag{10}$$

Next, we calculate the conformity scores $\hat{\varepsilon}_i$, which are trained on the calibration set, to evaluate the magnitude of the error relative to the lower and upper bounds of the interval. The scores are thus given by:

$$\hat{\varepsilon}_i = \ max\left\{\hat{q}_{\alpha_L}(\boldsymbol{X}_i) - Y_i, Y_i - \hat{q}_{\alpha_H}(\boldsymbol{X}_i)\right\},\ \forall\, i \in I_2. \tag{11}$$

Finally, after using Equation (8) to obtain $q_{1-\alpha}\,(\hat{\varepsilon}_i, I_2)$ from the empirical quantile distribution of $\hat{\varepsilon}_i :, i \in I_2,$ we can compute the conformalized quantile prediction interval of property prices given by:

$$C(\boldsymbol{X}_{n+1}) = \left[\hat{q}_{\alpha_L}(\boldsymbol{X}_{n+1}) - q_{1-\alpha}\,(\hat{\varepsilon}, I_2), \hat{q}_{\alpha_H}(\boldsymbol{X}_{n+1}) + q_{1-\alpha}\,(\hat{\varepsilon}, I_2)\,\right]. \tag{12}$$

The last step involves assessing the performance of the model by examining the empirical coverage properties of the conformalized quantile prediction interval obtained in Equation (12). The model is deemed to be adaptive if the prediction interval adjusts itself based on the characteristics of the observed data. Furthermore, it is considered well-calibrated if the empirical coverage aligns closely with the nominal coverage level (Angelopoulos & Bates, 2023).

The following theorem from Romano et al., (2019) provides validity for the conformal quantile procedure explained in this section.

Theorem Romano et al., (2019). If $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, the prediction interval $C(\boldsymbol{X}_{n+1})$ given in Equation (12) satisfies:

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

This is because exchangeability implies that the order in which the data points are observed does not affect their joint distribution. As such, under exchangeability, the joint distribution is consistent across permutations of the data, ensuring the validity of the coverage guarantee.

Also, the prediction interval is nearly perfectly calibrated if the conformity scores $\hat{\varepsilon}_i$ are almost surely distinct[2]:

$$\mathbb{P}\{(Y_{n+1} \in C(\boldsymbol{X}_{n+1})\} \geq 1 - \alpha\ + \frac{1}{1 + n_2}.$$

3.2 Gradient boosted tree for quantile

For the purpose of this dissertation, a gradient boosting (Freund & Schapire , 1996; Friedman, 2001) regression model is used to obtain the conformal prediction intervals of property prices. Gradient boosting machine is an ensemble learning method which combines multiple weak learners to build a strong predictive model (Hjort, et al., 2022). The model we employ differs slightly from a regular gradient boosting machine, as we are applying it to quantiles. In this

[2] The term "almost surely distinct" indicates that the conformity scores are distinct with probability one.

case, we use a pinball loss function, similar to the one defined in Equation (4), rather than the squared-error loss function typically used when boosting. By training K decision trees $\{ h_k(\boldsymbol{X})\}_{k=1}^{K}$ in a sequential manner, and summing the resulting predictions, we obtain an estimate for $\hat{Y}$ given by:

$$\hat{Y} = \sum_{k=1}^{K} h_k(\boldsymbol{X}). \tag{13}$$

The first decision tree, $h_1(\boldsymbol{X})$ is trained on the original data, and the subsequent trees are added to the ensemble. The gradient boosting machine iteratively trains the decision trees and, on each iteration, tries to correct the errors made by the previous ones. The algorithm will assign weights to the data points with larger residuals so that it can focus on quantiles that are harder to predict. During each iteration, the gradient boosting machine will calculate the gradient of the loss function with respect to the predicted quantiles. The pinball loss function, which the algorithm tries to minimize, is given by:

$$\rho_L(Y_i, \hat{Y}_i^{(k-1)} + h_k(\boldsymbol{X_i})) = \begin{cases} \alpha(Y_i - \hat{Y}_i^{(k-1)} - h_k(\boldsymbol{X_i})), & if\ Y_i \geq \hat{Y}_i^{(k-1)} + h_k(\boldsymbol{X_i}), \\ (1-\alpha)(\hat{Y}_i^{(k-1)} - h_k(\boldsymbol{X_i}) - Y_i) & otherwise. \end{cases} \tag{14}$$

The function will correct the underestimation and the overestimation of the quantiles through penalty terms from the regularized loss function:

$$\sum_{i=1}^{n} \rho_L(Y_i, \hat{Y}_i^{(k-1)} + h_k(\boldsymbol{X_i})) + \gamma T + \frac{1}{2}\eta||\boldsymbol{w}_k||^2. \tag{15}$$

The parameters γ and η penalize the number of terminal nodes and the magnitude of the weights, respectively. For optimization, this dissertation applies a gradient descent, namely the Light Gradient Boosting Machine (LightGBM), on the loss function. This method is chosen as it has the fastest computational speed compared to the other baseline algorithms, such as the widely known eXtreme Gradient Boosting (XGBoost), and it preserves its high accuracy level. Also, considering that this algorithm is histogram-based, it requires less memory consumption (Ke, et al., 2017).

4. EMPIRICAL APPLICATION

4.1 Data

The empirical analysis for this study is conducted using property prices of San Francisco's Bay Area, California USA. The residential price data consists of 8,351 observations on sold properties over the period 2020 to 2023. The dataset was obtained using an API service as large residential datasets on sold properties are limited and often privately owned by real estate companies. The United States Census Bureau website provides geographical data with a vast range of demographic attributes which allowed us to enrich the dataset for the analysis. The additional variables created based on the zip codes provided in the original dataset include, neighbourhood delimitations, median income per neighbourhood per year, population, race, poverty rate, employment rate, homeownership rate and the number of housing units available in each neighbourhood. Spatial heterogeneity—the varied distribution of characteristics across different geographic locations—plays an essential role in influencing property prices. As previously discussed in Section 2.1 of this dissertation, the quality of neighborhoods and economic dynamics deeply influence property values. Factors like safety, poverty rate, employment rate, and median household income contribute to the overall appeal of an area. Well-maintained, secure, and economically robust neighborhoods often attract larger population, leading to higher property prices as they are deemed more attractive to potential buyers. Given that spatial heterogeneity highly influences property prices, defining the demographic characteristics for all neighbourhoods in our dataset allowed us to yield more accurate predictions and thus increase the overall performance of the model.

As described in Table 6 found in Appendix A, the housing types included in the analysis vary from single-family homes, townhouse, condos, and apartments, with single-family homes representing most of the observations (approximately 92% of the data points). The average sold property in the dataset is a single-family 2004 square foot home, built mid-1930s, with 3-bedrooms, 2-bathrooms, a parking space, and central heating. Moreover, the range of sold property prices in the dataset is very large, with a minimum value equal to $50,000 and maximum of $43,500,000 – price expressed in USD. The observed residential prices in the San Francisco Bay Area are highly skewed to the left and average around $2,148,723, including all property types. Figure 1 illustrates the distribution of house prices and a subset of regressors. We observe an asymmetric distribution for those variables further confirming the positive skewness, which is expected due to the nature of the data.

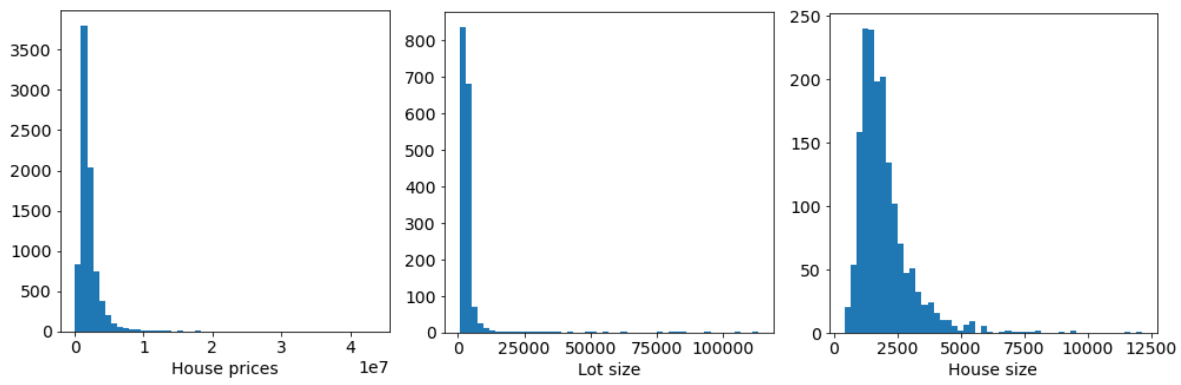


FIGURE 1: Histogram of a subset of the variables: house prices, lot size, house size, and number of bathrooms.

Furthermore, we observe that there is considerable variability in the data, both within individual variables—as evidenced by large and varying standard deviations in Table 5 found in Appendix A—and in the relationship between the price and different housing attributes, as illustrated in Figure 2. This variability points to complex dynamics between property price and the covariates in the dataset, and could potentially indicate heteroskedasticity in the data. However, this is not worrisome considering that conformal quantile regression is adaptive to heterogeneous data (Romano et al., 2019).

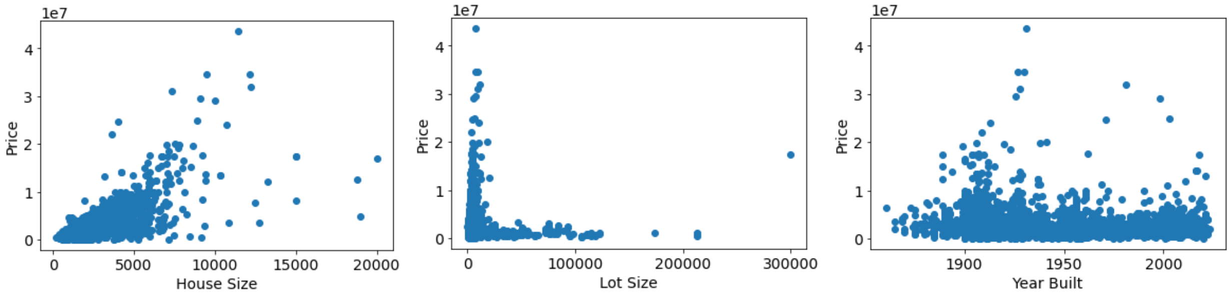


FIGURE 2: Scatter plots of property price against a subset of the variables: house size, lot size, and year built.

The median income in San Francisco is $131,288 per year and exhibits high income variation, as indicated in Table 5. Additionally, the relationship between different income levels in specific neighbourhoods and the average property price within those areas is illustrated in Table 1. We note that neighbourhoods within the lower income category, i.e., with a median household income below the San Francisco average, generally have the lowest median property values. Properties in middle- and upper-class neighbourhoods average around $2.1 million and $2.7 million, respectively. Hence, we find a positive correlation between income level and property value as higher median income brackets are associated with increasing average property price.

	Minimum (Maximum)	Neighbourhoods	Average Property Price
Lower Income	55,888 (93,995)	Van Ness/Civic Center, South of Market, Lower Nob Hill/Chinatown/Downtown, Marina District, Bayview-Hunters Point, Portola	1,203,041.79
Average Income[3]	104,476 (161,391)	Polk/Russian Hill (Nob Hill), Embarcadero, Inner Mission, Mission Terrace, Parkside/ Sunset District, Zion District/Lower Pacific Heights	2,147,108.2
High Income	164,289 (244,662)	Financial District South, Castro, Diamond Heights/Twin Peaks West, Marina, Mission Bay, Westwood Highlands/Twin Peaks West	2,669,020.8

TABLE 1: Neighborhoods of San Francisco categorized by various income levels and their corresponding average property price.

From the lower income bracket, the neighbourhood of Van Ness/ Civic Center reports the lowest median household income at $55,888. Additionally, this neighborhood faces a poverty rate of 18.5%, positioning it as the second most impoverished area in San Francisco, following Lower Nob Hill/Chinatown/Downtown. On the other hand, the Financial District South represents the neighbourhood with the highest annual median income of $244,662. Accordingly, the area is amongst the neighbourhoods with the highest property prices, as well as the highest employment and home ownership rates with 76% and 41.8%, respectively. The discrepancies in socio-economic indicators between the different areas in San Francisco underscores the spatial heterogeneity in the city. This further emphasizes the link between income levels, associated socio-economic indicators, and the necessity to include those types of variables in the property valuation process.

It's crucial to highlight that, despite identifying a positive correlation between income level and property value in specific neighbourhoods, significant price fluctuations are noticeable within each individual neighborhood. Zion District/Lower Pacific Heights, for instance, a neighbourhood where the median income aligns closely with the San Francisco

[3] The salaries included in the average income category are within approximately ±$30,000 from the median income in San Francisco of $131,288 per year.

average—roughly \$7,000 higher—exhibits considerable variation in property prices, ranging from a minimum of \$388,498 to a maximum of \$43,500,000. It is important to acknowledge these variations and especially the extreme values, as they can impact the ease of prediction and, consequently, the widths of the predicted confidence intervals.

4.2 Methodology

The steps we employ to construct the conformal intervals of property prices are the following.

1. Set the error rate α at 0.10, such that 90% of the actual property prices fall within the constructed conformal intervals.
2. Randomly split the dataset into a training set S_t and a calibration set S_c, which contain 80% and 20% of the observations, respectively.
3. Use the gradient boosting tree algorithm to fit the training set on two conditional quantiles $\{\hat{q}_{\alpha_L}, \hat{q}_{\alpha_H}\}$.
4. Evaluate, on the test data, the adaptivity of the model based on the empirical coverage properties.[4]

Quantiles are sensitive to hyperparameters – number of decision trees, depth of the tree, maximum number of terminal nodes in one tree, learning rate – and often yield intervals that are too wide (Romano, et al., 2019). To mitigate this issue, the following steps were employed.

a. We specified a range of possible values for the number of trees since, up to a turning point, increasing the number of trees has a diminishing marginal effect on the out-of-sample accuracy. We set the maximum number of trees in the ensemble to $\in \{500, 1000, 1500\}$.

b. We restricted the number of terminal nodes in one tree. This aims to control the complexity of the gradient boosting model and helps mitigate the risk of overfitting. A tree with too many leaves tends to follow the training data too closely, resulting in poor out-of-sample accuracy. The maximum number of leaves are $\in \{32, 64, 128, 256\}$.

c. We specified a maximum depth of the decision tree, which is expected to introduce bias in the model, but aims to reduce the variance. This bias-variance trade-off prevents

[4] The conditional coverage will also be evaluated although we understand that the guarantee of this coverage is only valid under strong distributional assumptions and thus is rarely satisfied.

overfitting and preserves out-of-sample accuracy. The maximum depth of the tree is set to $\in \{8, 16, 32, 64\}$.

d. We specified a range of values for the learning rate. This controls the convergence speed and determines the step size at which the model moves toward minimizing the residuals. A small learning rate will act as a regularization term trying to prevent the model from overfitting the data. The learning rate was set to $\in \{0.01, 0.05, 0.1\}$.

e. We used a tuning technique known as grid-search to determine all these optimal hyperparameters for the models.

The combination of hyperparameters that yields the lowest mean absolute error (MAE) on the validation data was selected to train the conformal quantile regression model.

4.3 Marginal coverage

Marginal coverage in the context of conformal prediction refers to the share of actual observations that fall inside the constructed prediction interval given a pre-defined error rate (Angelopoulos & Bates, 2023). By calculating the empirical coverage, we can evaluate whether the finite-sample marginal coverage guarantee given by Equation (5) holds. If the calculated coverage is significantly lower or higher, it may indicate an issue with the size of the calibration set. That is, since the calibration set changes at each iteration of the algorithm, the coverage of the conformal prediction is random. Given the randomness of the coverage obtained each time we sample a new calibration set, the following Beta distribution is introduced by Vovk (2012) to model the distribution of the coverage.

$$\mathbb{P}(Y_{test} \in C(Y_{test}) \mid \{(X_i, Y_i)\}_{i=1}^{n}) \sim Beta\ (n + 1 - m, m), \tag{16}$$

where $m = (n + 1)\alpha$. Hence, by setting the size of the calibration set sufficiently high, we can obtain more precise estimates of non-conformity scores thereby resulting in narrower prediction intervals and valid empirical coverage, centered around the desired $1 - \alpha$ (Angelopoulos & Bates, 2023). Also, by running the conformal prediction model multiple times with different training, calibration, and test sets we can capture the variation in the limited sample size and mitigate the issue of finite-sample variability. Given this, 100 random splits into training, calibration, and test sets were performed to obtain both the standard and conformal quantile regressions coverages and widths displayed in Table 2.

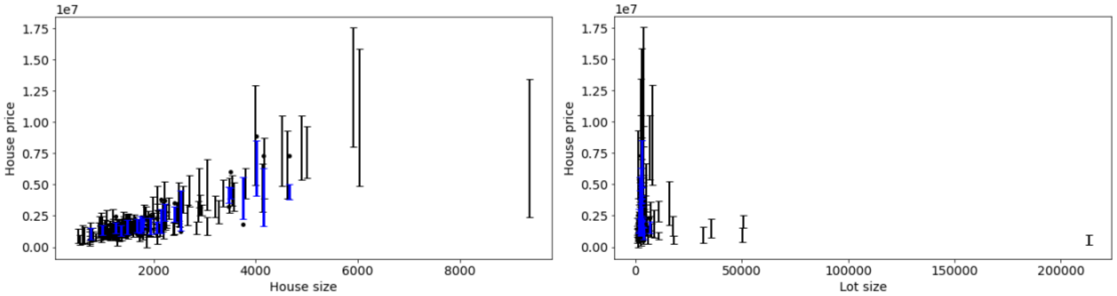
Relative median width of the prediction intervals is another measure we use to evaluate the adaptivity of the model. Given that our dataset contains multiple extreme values, which can cause the absolute width of the prediction intervals to be disproportionately large, we are using the median relative width as it is less sensitive to skewness in the data.

	Conformal quantile regression	Standard quantile regression
Marginal coverage	0.90	0.70
Median width	0.68	0.40

TABLE 2: Marginal coverage and relative median width of the prediction intervals for both the conformal quantile regression and the standard quantile regression. All figures are averages based on the 100 random splits of the data into training, calibration, and test sets.

The results in Table 2 illustrate that the conformal quantile regression provides a marginal coverage on the test data that is centered at the nominal coverage level set at 90%. The standard quantile regression on the other hand reports a marginal coverage that significantly deviates from the desired nominal level. This is expected considering that conformal prediction guarantees that, on average, the marginal coverage will be equal or exceed the desired nominal level, whereas standard quantile prediction does not have this theoretical guarantee.

Figure 3 illustrates the conformal prediction intervals of house prices against a subset of regressors from the dataset: house size (top left), lot size (top right), median income (bottom left), and housing units (bottom right). The black dots represent the true house prices in the test data, the black bars represent the cases where the predictions fall within the prediction intervals and the blue bars represent predictions that fall outside. The cases where the prediction interval fails to cover the true prices represent approximately the miscoverage rate set at 10%. We notice that as prices and house sizes get larger the conformal prediction intervals accordingly get larger. This suggests that the conformal model is adaptive to uncertainty and variation observed in the data.



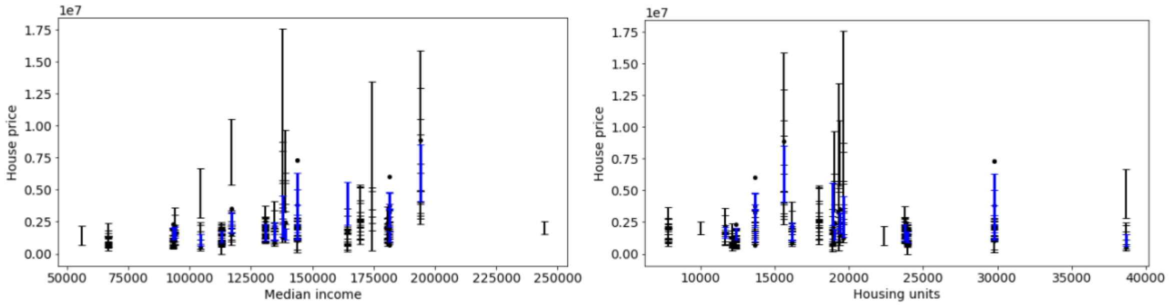


FIGURE 3: Conformal prediction intervals of house prices against a subset of regressors from the dataset: house size (top left), lot size (top right), median income (bottom left), and housing units (bottom right). The dots represent the actual house prices from the test data. The black bars represent the prediction inside the prediction intervals and the blue bars represent one outside. The nominal coverage is 0.90.

As for the relative median widths, the conformal values tend to be more conservative, in that they are generally larger than their non-conformal counterparts. This result comes from the fact that the width of conformal models is adaptive to variability in the data. Hence, intervals will widen according to the uncertainty associated with the predicted property prices to achieve adequate coverage. Although the non-conformal width is significantly lower than the conformal one, both regression methods yield relatively large widths, indicating uncertainty associated with the predicted intervals. However, the coverage rate of the conformal quantile regression maintains the desired nominal coverage level.

As aforementioned, the presence of extreme values limits the dispersion of property prices in the tails of the distribution. This limitation makes it difficult to accurately estimate quantiles, introducing greater uncertainty into the prediction intervals. Consequently, the insufficient data in the tails contributes to broader intervals, making the model more conservative to account for potential variability. Given that our dataset set contains a small number of very large extreme values which translates into large relative median widths, we decreased the nominal coverage level set initially. The aim is to obtain narrower prediction intervals, providing increased precision but at the potential cost of a higher risk of undercoverage. Setting the pre-defined error rate α to 0.2, we obtained a marginal coverage on the test data that is centered at the desired nominal level of $0.799 \cong 80\%$, and as expected the median relative width decreased to 0.485. While the width did decrease, it would be more satisfactory if the reduction was more substantial. Section 5 below provides a more detailed exploration of the limitations of this analysis, shedding light on other factors that might contribute to consistently wide widths.

4.4 Conditional coverage

Conditional coverages of conformal quantile predictions measure the reliability of prediction intervals conditional on a specific variable (Angelopoulos & Bates, 2023). As previously shown, Equation (1) must hold for any $\boldsymbol{X}_{n+1}$ to satisfy the conditional coverage guarantee. Considering the strong assumptions regarding the joint distribution $P_{\boldsymbol{X},Y}$ required for the conditional guarantee to hold, in practice, it rarely does (Sesia & Candes, 2020). We will nonetheless evaluate the conditional coverages to assess how the model behaves when applied to a subset of variables in the dataset.

Bin cuts

Variable	Lower	Middle	Upper
House size	< 1500	$1500 \leq HS < 2000$	≥ 2000
Lot size	< 3000	$3000 \leq LS < 6500$	≥ 6500
Median income	< 125000	$125000 \leq MI < 175000$	≥ 175000
Housing units	< 15000	$15000 \leq HU < 20000$	≥ 20000

	Conditional coverage			**Median width**		
Variable	Lower	Middle	Upper	Lower	Middle	Upper
House size	0.95	0.92	0.84	0.76	0.63	0.63
Lot size	0.91	0.88	0.90	0.68	0.66	0.94
Median income	0.94	0.88	0.88	0.74	0.64	0.62
Housing units	0.93	0.87	0.91	0.68	0.68	0.70

TABLE 3: The cut off values for each variable are given in the top part of the table. The test data was used to evaluate the bin cuts. The bottom part of the table reports the conditional coverages and median relative widths of the conformal prediction intervals. All figures are averages based on the 100 random splits of the data into training, calibration, and test sets.

To obtain the conditional coverages and median widths of different variables in the dataset we first evaluated the spread of test observations of each regressor. The bins were determined based on the clustering pattern observed in the test data, dividing the observations into lower, middle, and upper segments. After setting the appropriate bin cuts, the whole dataset set was used to train the model on a defined section of the regressor space. The top part of Table 3 shows the different cut levels per chosen regressor—house size, lot size, median income, and housing units—and the bottom part displays the resulting conditional coverages and relative median widths. Based on these results, we observe that for each category, the proportion of property prices that fall within the predicted interval is centered around the desired nominal level of 0.90. As for the median widths, we observe relatively varying widths for all four regressors and for each bin level. However, we notice that the widths are relatively large, which indicates lower reliability of the interval predictions. Although the median width is used to mitigate the effect of extreme values, the limited number of observations in the tail of the distribution of prices combined a relatively small and highly variable dataset leads to large interval widths. Again, this is required to account for the substantial variability observed in the data and ensure that the intervals cover a wide range of predicted property values.

As depicted in Table 4, we re-evaluated the model, using the same bin cuts, but reduced the nominal coverage level to 80%. This modification enabled us to observe the behavior of the median relative widths with respect to specific covariates. As expected, the widths for all examined regressors decrease significantly, suggesting improved precision. Notably, the resulting conditional coverage levels stand around the desired level 80%, and in certain instances, they exceed this threshold, reaching close to 90%. This suggests the model is exhibiting favorable behavior, providing narrower prediction intervals while maintaining robust coverage levels.

	Conditional coverage			**Median width**		
Variable	Lower	Middle	Upper	Lower	Middle	Upper
House size	0.87	0.82	0.72	0.54	0.45	0.46
Lot size	0.82	0.77	0.82	0.49	0.48	0.66
Median income	0.86	0.77	0.77	0.53	0.46	0.45
Housing units	0.84	0.78	0.81	0.49	0.49	0.51

TABLE 4: The cut off values for each variable are given in the top part of Table 3. The table reports the conditional coverages and median relative widths of the conformal prediction intervals given a confidence level set to 0.80. All figures are averages based on the 100 random splits of the data into training, calibration, and test sets.

5. LIMITATIONS

The widths of prediction intervals in conformal quantile regression are influenced by several factors and in the context of this dissertation, the following issues were identified as possible cause for large prediction interval widths. Firstly, the empirical analysis relies on a relatively small dataset. In cases of insufficient data, quantile estimates become imprecise, resulting in increased uncertainty and wider prediction intervals. Secondly, the presence of a small number of very large extreme values led the model to adopt a conservative approach to prevent undercoverage. This in turn resulted in larger intervals to ensure adequate coverage of the true response. The third factor concerns the limited inclusion of relevant explanatory variables. The accuracy of property price predictions depends on the quality and quantity of housing attributes considered. Insufficient relevant variables hinder on the model's ability to comprehensively explain data variability, contributing to increased uncertainty and wider prediction intervals. Future research in this area should consider a larger dataset and an expanded set of factors influencing price variation, such as location-specific amenities and services (e.g., proximity to schools, parks, public transportation, hospitals). These enhancements should help increase model precision and narrow prediction intervals.

6. CONCLUSION

In this dissertation, a real estate appraisal model based on a gradient boosted tree for quantiles was applied. The findings of this analysis show that by employing the model-agnostic technique of conformal prediction to quantile regression we may quantify uncertainty associated with the prediction of housing prices. This method allowed us to construct reliable prediction regions without distributional assumptions, thereby guaranteeing valid coverage in finite-samples. Through an empirical application on property prices of the San Francisco Bay Area, we observed that conformal quantile prediction, consistently delivers coverage guarantees centered around the desired nominal level. Also, we showed that the intervals are adaptive to the variations in the data, which is crucial considering the heterogeneous nature of real estate data. The ability to flexibly adjust the width of prediction intervals to accommodate varying levels of uncertainty and complexity within the dataset is a significant advantage. It enables real estate professionals to make informed decisions across a wide spectrum of property types, locations, and market conditions, thereby enhancing the practical utility of the method.

REFERENCES

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

Babu, A. & Chandran, A. 2019. Literature review on real estate value prediction using machine learning. *International Journal of Computer Science and Mobile Applications, 7*(3), pp. 8-15.

Bai, Y., Mei, S., Wang, H. & Xiong, C. 2021. Understanding the under-coverage bias in uncertainty estimation. ArXiv, abs/2106.05515.

Bellotti, A. 2017. Reliable region predictions for automated valuation models. *Annals of Mathematics and Artificial Intelligence, 81*, pp. 71-84.

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

Can, A. 1992. Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics, 22*, pp. 435-474.

Chau, K. W. & Chin, T. L. 2003. A critical review of literature on the hedonic price model. *International Journal for Housing Science and Its Application, 27*(2), pp. 145-165.

Court, A. T. 1939. Hedonic price indexes with automotive examples. *The Dynamics of Automobile Demand,* pp.99-119.

Feldman, S., Bates, S. & Romano, Y. 2021. Improving conditional coverage via orthogonal quantile regression. *Advances in Neural Information Processing Systems.*

Freund, Y. & Schapire , R. E. 1996. Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning,* pp. 148-156.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), pp. 1189-1232.

Goodman, A. C. 1978. Hedonic prices, price indices and housing markets. *Journal of Urban Economics, 5*, pp. 471-484.

Griliches, Z. 1971. Introduction: Hedonic price indexes revisited. In: Griliches, Z. ed. *Price Indexes and Quality Change: Studies in New Methods of Measurement*. Cambridge, MA and London, England: Harvard University Press, pp. 3-15.

Grudnitski, G. & Do, A. Q. 1992. A neural network approach to residentia property appraisal. *Real Estate Appraiser, 58*(3), pp. 38-45.

Hjort, A., Pensar, J., Scheel, I. & Sommervoll, D. E. 2022. House price prediction with gradient boosted trees under loss functions. *Journal of Property Research, 39*(4), pp. 338-364.

Ho, W. K., Tang, B.-S. & Wong, S. W. 2021. Predicting property prices with machine learning algorithms. *Journal of Property Research, 38*(1), pp. 48-70.

Hu, Y., Musielewicz, J., Ulissi, Z. W. & Medford, A. J. 2022. Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Machine Learning: Science and Technology, 3.*

Ke, G. et al. 2017. LightGBM: A highly effecient gradient boosting decision tree. *Advances in Neural Information Processing Systems.*

Kecman, V. 2005. Support vector machines: An introdution. In: Wang, L.P., Ed., *Support Vector Machines: Theory and Applications*, Springer, Berlin, pp. 1-47.

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

Limsombunchai, V., Lee, M., Gan, C. & Good, A. 2004. House price prediction: Hedonic price model vs. artificial neural network. *American Journal of Applied Sciences, 1*(3), pp. 193-201.

McCluskey, W. et al. 2012. The potential of artificial neural networks in mass appraisal: the case revisited. *Journal of Financial Management of Property and Construction, 17*(3), pp. 274-292.

Papadopoulos, H., Proedrou, K., Vovk, V. & Grammerman, A. 2002. Inductive confidence machines for regression. *Proceedings of Machine Learning: European Conference on Machine Learning*, pp. 345-356.

Peter, N. J., Okagbue, H. I., Obasi, E. C. & Akinola, A. O. 2020. Review on the application of artificial neural networks in real estate valuation. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(3), pp. 2918-2925.

Romano, Y., Patterson, E. & Candes, E. J. 2019. Conformalized quantile regression. *Advances in Neural Information Processing Systems, 32*, pp. 3543–3553.

Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), pp. 34-55.

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

Vovk, V. 2012. Conditional validity of inductive conformal predictors. *Proceedings of Machine Learning: Asian Conference on Machine Learning* 25, pp. 475-490.

Wang, C. & Wu, H. 2018. A new machine learning approach to house price estimation. *New Trends in Mathematical Sciences, 6*(4), pp. 165-171.

White, H. 1989. Learning in artifical neural networks: A statistical perspective. *Neural Computation, 1*(4), pp. 425-464.

APPENDICES

A. Summary statistics

TABLE 5 Shows the summary statistics of the numerical variables in the dataset.

Variable	Unit	Mean (Std. Deviation)	Minimum (Maximum)	Type
Property Price	USD	2148723 (1967659)	50000 (43500000)	Numerical
House Size	Square footage	2003.989 (1171.278)	200 (20000)	Numerical
Lot Size	Square footage	4207.132 (10652.23)	100 (299475)	Numerical
Bedrooms	–	3.164172 (1.155457)	1 (15)	Numerical
Bathrooms	–	2.435996 (1.292704)	1 (13)	Numerical
Year Built	Years	1936.223 (28.84216)	1861 (2023)	Numerical
Parking	–	1.276494 (8.8427625)	0 (5)	Numerical
Median Income per Neighbourhood	USD	136935.1 (35244.85)	55888 (244662)	Numerical
Population	–	44222.6 (17197.83)	4306 (79314)	Numerical
Employment Rate	Percentage	66.59686 (6.227736)	54.9 (79.6)	Numerical
Poverty Rate	Percentage	8.643995 (3.546486)	4.4 (19.7)	Numerical
Homeownership Rate	Percentage	51.83847 (16.89804)	9.2 (81.1)	Numerical
Housing Units	–	18142.46 (6415.972)	2906 (38664)	Numerical
Black/ African American	–	1945.309 (2175.857)	124 (10143)	Numerical
Asian	–	16081.24 (10676.85)	1839 (40116)	Numerical
Hispanic/ Latino	–	7700.451 (7077.502)	293 (22160)	Numerical
White	–	16898.91 (7464.099)	1902 (31306)	Numerical

Table 6 Shows the summary statistics of the categorical variables in the dataset.

Variable		Frequency (Percentage)	Cumulative Frequency	Type
House Type	Single family	7721 (92.46)	92.46	Categorical
	Townhouse	91 (1.09)	93.55	
	Condo	518 (6.20)	99.75	
	Apartment	21 (0.25)	100.00	
Heating	Electric	408 (4.89)	4.89	Categorical
	Central	7051 (84.43)	89.32	
	Radiant	723 (8.66)	97.98	
	No heating	169 (2.02)	100.00	
Fireplace	No fireplace	7796 (93.35)	93.35	Binary
	Fireplace	555 (6.65)	100	
Neighbourhood	Van Ness/Civic Center	48 (0.57)	0.57	Categorical
	South of Market	30 (0.36)	0.93	
	Financial District South	39 (0.47)	1.40	
	Mission Bay	250 (0.22)	4.39	
	Lower Nob Hill/Chinatown/Downtown	18 (2.02)	4.61	
	Polk/Russian Hill (Nob Hill)	121 (1.45)	6.06	
	Inner Mission	682 (8.17)	14.23	
	Embarcadero	9 (0.11)	14.33	
	Mission Terrace	816 (9.77)	24.10	
	Castro	529 (6.33)	30.44	

Zion District/Lower Pacific Heights	234 (2.80)	33.24
Parkside/ Sunset District	798 (9.56)	42.80
Buena Vista Park	248 (2.97)	45.77
Inner Richmond/Richemond District	346 (4.14)	49.91
Outer Richmond	471 (5.64)	55.55
Marina	256 (3.07)	58.62
Bayview-Hunters Point	451 (5.40)	64.02
Westwood Highlands/Twin Peaks West	666 (7.98)	71.99
Diamond Heights/Twin Peaks West	695 (8.32)	80.31
Stonestown	353 (4.23)	84.54
Marina District	21 (0.25)	84.79
Portola	543 (6.50)	91.29
Central Sunset/ Sunset District	727 (8.71)	100.00