Benchmarking Advanced Machine Learning for Accurate Property Valuation: A Replication Study

Name: Prince Djanku| |**Program:** Global Management| |**Date:** 03 Nov 2024

Abstract

The study presented here aims to enhance the accuracy of property price predictions for residential properties in King County, Washington, by benchmarking advanced machine learning algorithms. The research utilized a dataset of 21,613 property sales and incorporated 19 predictive variables to evaluate the performance of Multiple Linear Regression, Ridge Regression, Lasso Regression, Random Forest, and Cubic Polynomial Regression.

The findings of this study indicate that the Random Forest model outperformed the other techniques, achieving the lowest mean squared error (MSE) of 0.0952, demonstrating its robustness in handling the complex, non-linear relationships within the dataset. In contrast, the Ridge and Lasso regression models provided minor improvements in coefficient stability but did not surpass the predictive accuracy of linear regression. The polynomial model, optimized at degree three, exhibited slight gains over linear models but introduced risks of overfitting at higher degrees.

These results corroborate prior studies highlighting the effectiveness of ensemble models particularly, Random Forest, in property valuation, underscoring their potential for reliable and scalable property appraisal. The insights gained from this research contribute to improved model selection strategies for property valuation, with implications for both homebuyers and financial institutions.

Keywords: House price, Machine learning algorithms, Evaluation metrics

1. Introduction

Accurate property valuation is of critical importance, as it underpins key financial decisions for both homebuyers and owners. However, as highlighted by Wilson et al. (2002) the popular "open market value" approach to property valuation is often unable to provide a reliable, sustainable assessment. As a result, inaccurate valuations have had significant social and economic consequences, such as negative equity, reduced labor mobility, and instability in the housing market (Daly et al. 2003; Matysiak et al.1995).

This study compares advanced machine learning methods for predicting house prices in Kings County, Washington, expanding on Ho et al.'s (2020) international peer-reviewed article titled "Predicting property prices with machine learning algorithms" published in the Journal of Property Research which Journal has a scientific level score of 1 in the field of business and finance by the Norwegian list (Kanalregister, n.d.). This high-ranking, peer-reviewed publication provides a strong foundation for this replication study. While Random Forest showed promising results, the authors acknowledged the need to benchmark its performance against other advanced machine learning techniques by incorporating more property features. The target paper analyzed 40,000 transactions with four property attributes in Hong Kong. However, this study will analyze 21,613 records with nineteen variables from the United States.

This paper will first provide a review of the relevant machine learning theories and concepts, including ensemble and linear methods, regularization, and feature engineering. The data collection and preprocessing procedures will then be detailed, followed by the model development and tuning process. The comparative analysis of the machine learning algorithms will be presented, with a focus on their predictive accuracy.

Importantly, this study will seek to verify the conclusions drawn in the prior work by (Ho et al., 2020). While their research demonstrated the potential of Random Forest regression for accurate property price prediction, the current analysis will determine whether their findings hold true when comparing Random Forest to other machine learning techniques, such as Multiple linear regression, Ridge regression and Lasso. Thus, by conducting a rigorous, comparative evaluation across multiple advanced algorithms, this research will provide a more comprehensive assessment of the optimal modeling approach for accurate and reliable residential property valuation in King County. The insights gained can then be used to inform and improve upon existing practices, ultimately benefiting both homebuyers and lenders in the local real estate market.

2. Literature review

The use of advanced machine learning techniques for accurate property price prediction has been a growing area of focus in real estate research (Deppner et al., 2023; Mullainathan & Spiess, 2017; Newell, 2024). Aligned with this Ho et al. (2021) compared Gradient Boosting Machine(GBM), Random Forest(RF), and Support Vector Machines (SVM) in their ability to predict housing prices using a large dataset of 40,000 transactions over 18 years in China with four features used in model developments. Their findings indicate that both RF significantly outperformed SVM and GBM, especially in datasets with highly complex interactions between variables. This finding is echoed

in the work of Bastos & Paquette (2024), who also found that ensemble methods like GBM yield more accurate predictions when applied to real estate data in the San Francisco Bay Area.

Traditional models often neglect the uncertainty inherent in these predictions, leading to potentially misleading appraisals. Bastos & Paquette (2024) address this gap by introducing conformal prediction methods to construct valid prediction intervals, which adapt to the heterogeneity of property data. Their study shows that larger and older properties present higher uncertainty in price predictions, a finding that emphasizes the need for more flexible machine learning models. Aligned with this, Hjort et al. (2022) noted that performance can vary significantly across different price segments, suggesting that models should be tailored to account for this variation.

The heterogeneity of real estate data, influenced by factors such as property size, age, and location, often challenges conventional models. For instance, Hjort et al. (2022) explored the impact of different loss functions in Gradient Boosted Trees and found that tailored loss functions improved predictive performance in lower-priced segments. This approach contrasts with the more standard use of squared error loss, which may not capture the nuances in lower and higher price ranges.

Building on these findings, researchers have begun to explore hybrid models that combine the strengths of different machine learning techniques. For example, Zhang et al. (2023) proposed a novel approach that integrates deep learning with traditional ensemble methods, demonstrating improved accuracy in predicting property prices across diverse market conditions. This hybrid approach not only captures complex non-linear relationships but also maintains the interpretability often associated with ensemble methods, addressing a key concern in the practical application of AI in real estate valuation. Furthermore, recent advancements in federated learning techniques offer promising solutions to the privacy concerns often associated with large-scale property data collection and analysis (Li et al., 2025).

While advanced machine learning techniques have shown promise in property price prediction, it is important to consider potential limitations and challenges. One significant counterargument is the issue of model interpretability and transparency. Complex machine learning models, particularly deep learning and ensemble methods, often function as "black boxes," making it difficult for stakeholders to understand how specific predictions are derived (Rudin, 2019). The lack of transparency in complex machine learning models for real estate pricing can be problematic due to financial and legal implications. These models may struggle to provide clear explanations for valuation decisions, potentially perpetuate biases from historical data, and face challenges in unprecedented market conditions. Additionally, the computational resources and expertise required may be prohibitive for smaller firms, potentially widening the gap between market participants. A balanced approach combining machine learning with traditional appraisal methods is necessary to ensure accuracy and transparency (Kleinberg et al., 2018; Del Giudice et al., 2020).

3. Methodology

Regression analysis is a statistical technique for understanding relationships between variables and predicting the value of a dependent variable based on one or more independent variables. It is used across various fields such as economics, biology, and engineering for informed decision-making and forecasting. The simplest form is linear regression, which assumes a linear relationship and represents it as a straight line on a graph. More complex forms like polynomial and logistic regression model non-linear relationships, providing greater accuracy for intricate data patterns. Polynomial regression captures curved relationships, while logistic regression predicts binary outcomes. Advanced techniques like machine learning algorithms, including random forests and neural networks, handle high-dimensional data and uncover hidden patterns that traditional methods may miss. We will now discuss specific regression analyses considered in this study but before we do so it is imperative to understand the goals of our model developments.

3.1 Bias and Variance paradox

The bias function represents the difference between the expected prediction of the model and the true value of the target variable. It can be expressed as Bias(x) = E[$\hat{f}$(x)] - f(x), where E[$\hat{f}$(x)] is the expected value of the predicted output, and f(x) is the true value of the target variable. This bias function measures the systematic error of the regression function, and a lower bias indicates a better fit to the true underlying relationship (Glauner et al., 2018). However, the variance function, on the other hand, represents the variability of the model's predictions across different subsets of the data (Tibshirani & Tibshirani, 2009). Rasmussen and Williams (2005) expressed it as Var(x) = E[($\hat{f}$(x) - E[$\hat{f}$(x)])^2], where E[$\hat{f}$(x)] is the expected value of the predicted output, and $\hat{f}$(x) is the predicted output and $\hat{f}$(x) is the predicted output of the regression function. Thus, the variance function measures the sensitivity of the regression function to the specific training data used, and a lower variance indicates a more stable and consistent model.

Therefore, the goal in any regression problem is to find a model that minimizes both the bias and variance, as this will result in a more accurate and generalized regression function. Techniques such as cross-validation can help achieve this balance. Thus, by understanding and managing the bias and variance functions, we can develop more robust and reliable regression models that are well-suited for predicting property prices.

3.2 Multiple linear regression

Multiple regression is a powerful statistical technique that allows researchers to model the relationship between a single dependent variable and multiple independent variables simultaneously (Lee et al., 2019). This approach extends the simple linear regression model by incorporating additional predictors, providing a more comprehensive understanding of the factors influencing the dependent variable. This approach is ideal for modeling data where the predictor is a continuous dependent variable (Lee, 2022).

General form: (Hastie et al., 2001)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Where:
Y = dependent variable
$X_1, X_2, ..., X_p$ = independent variables
β_0 = y-intercept (constant term)
$\beta_1, \beta_2, ..., \beta_p$ = regression coefficients
ε = error term

The dependent variable Y represents the outcome that we want to model. In other words, the variable we want to find (predict). The independent variable ($X_1, X_2, ..., X_p$) are the variables that you believe influence or affect the dependent variable (Y). The constant term β0 represents the starting point of the regression line. Thus, the value of Y when all the independent variables are equal to zero. The regression coefficients ($\beta_1, \beta_2, ..., \beta_p$) represent the expected change in the dependent variable (Y) when the corresponding independent variable ($X_1, X_2, ..., X_p$) increases by one unit, while holding all other variables constant. Thus, when all other variables equal zero. The regression coefficients also indicate the strength and direction (negative or positive) of the relationship between each independent variable and the dependent variable. Lastly the error term epsilon denoted by (ε) captures the effects of other factors or variables that may have influence on the predictor but have not been accounted for in the model as well as any random errors or noise in the data that is used for the prediction. The goal is to find the best-fitting linear equation that describes the relationship between the dependent variable (Y) and the multiple independent variables ($X_1, X_2, ..., X_p$).

The most common method to estimate the regression coefficients ($\beta_1, \beta_2, ..., \beta_p$) is Ordinary Least Squares (Myers et al., 2010). This approach minimizes the sum of squared residuals between the observed and predicted values (Bayen & Siauw, 2014). The resulting regression equation can then be used to make predictions about the dependent variable (Y) based on the values of the independent variables ($X_1, X_2, ..., X_p$). The strength and significance of the relationships between the independent variables and the dependent variable can also be assessed using statistical measures, such as the coefficient of determination (R-squared) and the p-values of the regression coefficients (Dionísio et al., 2006). In the case of a multiple linear regression AIC and BIC can also be used to compare the performance of different models with varying numbers of independent variables. Here the model with the lowest AIC or BIC value is generally the preferred model. Mean Squared Error (MSE) is also commonly used as a measure of accuracy or the goodness of fit in multiple linear regression models. In the context of multiple linear regression, the MSE is calculated as the average of the squared differences between the predicted values (from the regression model) and the actual observed values of the dependent variable (Wallach & Goffinet, 1989).

3.3 Ridge regression

One of the major assumptions of linear regression is that the variables in the dataset are not highly correlated. However, when the predictor variables are highly correlated, a phenomenon known as

multicollinearity can arise, leading to unreliable and unstable estimates of regression coefficients in ordinary least squares (OLS) regression. This is where Ridge Regression, a powerful technique developed by Hoerl and Kennard (1981) comes into play. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to disentangle the individual effects of the predictors on the outcome variable (Daoud, 2017).

Ridge Regression seeks to minimize the OLS function:

$$RSS = \Sigma (Yi - \hat{Y}i)^2 + \lambda \Sigma \beta j^2$$

Where:

$\Sigma (Yi - \hat{Y}i)^2$ is the sum of squared residuals (OLS objective) or MSEs

λ is the regularization parameter (tuning parameter)

$\Sigma \beta j^2$ is the sum of squared coefficients (excluding intercept)

The key advantage of Ridge Regression is its ability to address the issue of multicollinearity (Schreiber-Gregory, 2018). By introducing the shrinkage penalty, Ridge Regression shrinks the regression coefficients towards zero, reducing the impact of highly correlated predictors and leading to more stable and reliable estimates. This is particularly useful when dealing with high-dimensional datasets, where the number of predictors may exceed the number of observations making ridge regression a popular method in medical sciences.

Another important aspect of Ridge Regression is the selection of the tuning parameter, also called Lambda (λ). The value of λ determines the relative importance of the two terms in the objective function, and it is critical to choose the right value to achieve the optimal balance between model fit and coefficient shrinkage. Cross-validation is a commonly used technique for selecting the optimal value of λ.

It is worth noting that Ridge Regression has some limitations. Unlike other regression techniques, such as Lasso regression, Ridge Regression does not perform variable selection, as it does not force the coefficients to be exactly zero (Hastie, 2020). This can make the interpretation of the model more challenging, especially when dealing with a large number of predictors. Additionally, the standardization of the predictor variables is crucial in Ridge Regression, as the coefficient estimates can be sensitive to the scaling of the variables (García et al., 2015).

The key insight behind ridge regression is that by introducing a penalty term ($\lambda \Sigma \beta j^2$) that shrinks the magnitude of the regression coefficients, we can reduce the variance of the estimates at the expense of introducing some bias. If the decrease in variance is greater than the increase in squared bias, then the overall MSE of the model will decrease, leading to improved predictive performance compared to OLS.

3.4 Lasso Regression

The seminal work on Lasso Regression was conducted by Tibshirani, who introduced the method in 1996 (Tibshirani, 1996). Lasso Regression, or Least Absolute Shrinkage and Selection Operator,

is a type of linear regression that uses an L1 penalty to perform feature selection and regularization (Emmert-Streib & Dehmer, 2019). Unlike Ridge Regression, which uses an L2 penalty, Lasso Regression can shrink some coefficients to exactly zero, effectively removing irrelevant features from the model. This property of Lasso Regression makes it a useful tool for high-dimensional datasets, where the number of features may be much larger than the number of observations. Recent research has introduced several extensions to the basic Lasso Regression model, aiming to address its limitations (Zou & Hastie, 2005).

The Lasso Regression cost function is defined as:

$$\text{Loss} = \Sigma(Yi - \hat{Y}i)^2 + \lambda \Sigma|\beta j|$$

Where $\Sigma(Yi - \hat{Y}i)^2$ represents the mean squared error(MSE), βj are the model coefficients, and λ is the regularization parameter that controls the strength of the penalty.

The L1 penalty in Lasso is very greedy and as a result causes some coefficients to be shrunk to exactly zero, resulting in a sparse model that only includes the most important predictors. This makes Lasso Regression useful for feature selection, as it can automatically identify and exclude irrelevant variables from the model. Compared to Ridge Regression, Lasso Regression tends to handle high-dimensional data better, as it can simplify the model by excluding unimportant features. However, it may introduce more bias into the model due to the aggressive feature selection (Lemhadri et al., 2019).

3.5 Random Forest

Random Forests is an ensemble learning algorithm that builds upon the concept of bagging decision trees. It provides an improvement over standard bagged trees by introducing a small but clever tweak that helps decorrelate the individual trees in the ensemble (Svetnik et al., 2003). As in bagging, Random Forests constructs multiple decision trees, each trained on a bootstrapped sample of the original training data. However, the key difference is in how the trees are grown.

When building each tree in the Random Forests ensemble, at every split, the algorithm randomly selects a subset from a set of predictors selected from the total number of predictors for each split, rather than the full set (Svetnik et al., 2003). The split is then allowed to use only one of those subsets predictors. This random subset of predictors is chosen afresh for each split in the tree. Typically, the value of m is set to be around the square root of the total number of predictors.

Decorrelation function:

$$m \approx \sqrt{p}$$

where:

m represents the set of all subsets of predictors selected

p is the total number of predictors in the dataset

The rationale behind this approach is to prevent the trees from becoming too correlated with each other. If there is one very strong predictor in the dataset, a standard bagged tree ensemble would likely use this predictor at the top split of most trees, leading to highly similar trees and less reduction in variance through averaging (Mentch & Zhou, 2019). By restricting the predictor set at each split, Random Forests forces the trees to use different predictors, decorrelating the ensemble. This decorrelation is the key to the superior performance of Random Forests compared to bagging. When the individual trees are less correlated, averaging their predictions leads to a larger reduction in variance.

However, some argue that the decorrelation introduced by Random Forests is not always the key to its superior performance compared to bagging. While less correlated trees can lead to a larger reduction in variance, the specific problem and dataset at hand may benefit more from the increased model complexity and flexibility provided by the full set of predictors (Mentch & Zhou, 2019). In certain scenarios, the loss of predictive power from restricting the feature set at each split could outweigh the gains from decorrelation.

3.6 Polynomial regression

Polynomial models offer a powerful approach in machine learning for capturing non-linear relationships between input features and target variables. Unlike their linear counterparts, which assume a strictly linear dependence, polynomial models accommodate more intricate, curvilinear patterns often present in real-world data (Agarwal et al., 2014; Bai et al., 2009). This flexibility stems from incorporating higher-order terms of the input features, effectively allowing the model to learn a curved relationship.

The core of a polynomial model is represented by the equation

$$(y) = (\beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n)$$

where 'y' denotes the target variable, 'x' represents the input feature, ' β_0' is the intercept, and coefficients '$\beta_1 X$' through '$\beta_n X^n$' correspond to the weights assigned to polynomial terms up to the nth degree. Pasandi et al. (2020) explained that the degree ('n') acts as a tuning knob for model complexity. Higher degrees allow the model to learn highly non-linear relationships but can lead to overfitting, especially when data is limited. This occurs when the model learns the training data too well, capturing even the noise, and consequently fails to generalize to new, unseen data. Regularization techniques, such as Ridge Regression and Lasso Regression, address this by adding a penalty term to the model's cost function. This penalty discourages excessively large coefficients, effectively preventing the model from becoming overly complex and reducing overfitting. Choosing the right degree for a polynomial model is crucial. To mitigate this, techniques like regularization and cross-validation are essential for finding the optimal model complexity and ensuring the model generalizes well to unseen data (Bo et al., 2021). Metrics like mean squared error or R-squared can then be used to compare the performance of models with varying degrees and select the one that strikes a balance between bias and variance (Raschka, 2018).

4. Data

Ho et al. (2020) used machine learning algorithms to predict property prices based on data from 14 residential estates in a district in Hong Kong. The time series data that they used in their study span over 18 years and had about 40,000 records. However, the empirical results in this study are based on house prices from house sale prices for King County, which includes Seattle, USA. The dataset comprises 21,613 observations(houses) and 21 variables (house attributes) collected between May 2014 and May 2015 (Harlfoxem, 2016). This dataset was sourced from Kaggle.com, an online community platform for data scientists owned by google.

At the data preprocessing stage, we dropped the transaction id and transaction date columns from the dataset because these columns are not considered useful features in estimating property prices which brings the total number of variables under consideration to nineteen. Descriptive statistics was employed in python to understand the basic features and the measures in the dataset. Table 1 shows summary statistics for the variables in the data set. Key variables in the dataset include price, which is the target variable representing the sale price of homes. Independent variables such as the number of bedrooms, bathrooms, square footage of living space, and lot size serve as important predictors of house prices.

Table. 1 Summary statistics

	price	bedrooms	bathrooms	sqft_living	sqft_lot\
count	2.161300e+04	21613.000000	21613.000000	21613.000000	2.161300e+04
mean	5.400881e+05	3.370842	2.114757	2079.899736	1.510697e+04
std	3.671272e+05	0.930062	0.770163	918.440897	4.142051e+04
min	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02
25%	3.219500e+05	3.000000	1.750000	1427.000000	5.040000e+03
50%	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03
75%	6.450000e+05	4.000000	2.500000	2550.000000	1.068800e+04
max	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06

	floors	waterfront	view	condition	grade \
count	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
mean	1.494309	0.007542	0.234303	3.409430	7.656873
std	0.539989	0.086517	0.766318	0.650743	1.175459
min	1.000000	0.000000	0.000000	1.000000	1.000000
25%	1.000000	0.000000	0.000000	3.000000	7.000000
50%	1.500000	0.000000	0.000000	3.000000	7.000000
75%	2.000000	0.000000	0.000000	4.000000	8.000000
max	3.500000	1.000000	4.000000	5.000000	13.000000

	sqft_above	sqft_basement	yr_built	yr_renovated	zip code\
count	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
mean	1788.390691	291.509045	1971.005136	84.402258	98077.939805
std	828.090978	442.575043	29.373411	401.679240	53.505026
min	290.000000	0.000000	1900.000000	0.000000	98001.000000
25%	1190.000000	0.000000	1951.000000	0.000000	98033.000000
50%	1560.000000	0.000000	1975.000000	0.000000	98065.000000
75%	2210.000000	560.000000	1997.000000	0.000000	98118.000000

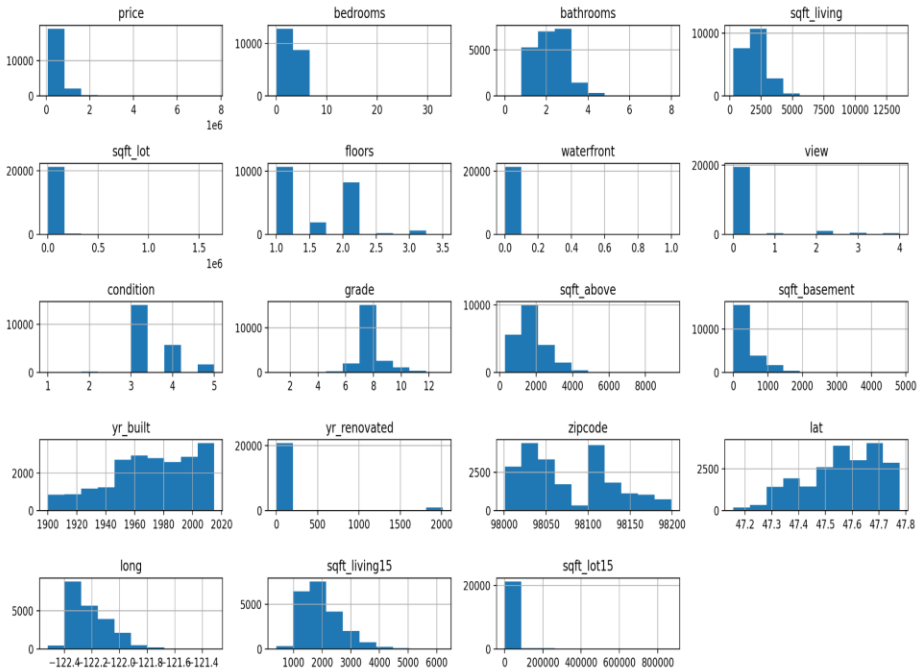
```
max     9410.000000    4820.000000   2015.000000   2015.000000  98199.000000

            lat          long        sqft_living15     sqft_lot15
count  21613.000000  21613.000000   21613.000000   21613.000000
mean      47.560053   -122.213896    1986.552492   12768.455652
std        0.138564      0.140828     685.391304   27304.179631
min       47.155900   -122.519000     399.000000     651.000000
25%       47.471000   -122.328000    1490.000000    5100.000000
50%       47.571800   -122.230000    1840.000000    7620.000000
75%       47.678000   -122.125000    2360.000000   10083.000000
max       47.777600   -121.315000    6210.000000  871200.000000
```

A careful analysis of the summary statistics reveals many notable characteristics in the dataset. For instance, the average price of the houses is approximately $540,088, accompanied by a substantial standard deviation of $367,127. This indicates a wide variation in property prices, ranging from a minimum of $75,000 to a maximum of $7,700,000. Such a broad price spectrum suggests the presence of both affordable housing options and luxury properties within the dataset. In terms of the number of bedrooms, the average is about 3, with a maximum of 33 bedrooms, which points to some outliers in the data. Meanwhile, the average number of bathrooms is approximately 2, with a maximum of 8 bathrooms. These findings imply that while most properties are likely to be standard family homes, there are also larger, possibly multi-family or luxury homes represented. The average living area of the properties is around 2,080 square feet square feet, with a maximum of 13,540 square feet, indicating a significant range in property sizes. Similarly, the average lot size is about 15,107 square feet, with a maximum of 1,651,359 square feet, suggesting that some properties may have extensive outdoor spaces.

These findings have important implications for the development and evaluation of predictive models. The wide range of prices and property characteristics suggests that the data is more suitable to robust algorithms capable of handling outliers and diverse data distributions. Models such as regression analysis may need to incorporate regularization techniques to mitigate the influence of extreme values.

The dataset also includes more nuanced features, such as the grade and condition of the houses, which capture the quality and upkeep of the properties, respectively. Geographical variables, such as the latitude and longitude, waterfront and non-waterfront allow for consideration of location-specific effects, which are critical in property valuation. Importantly, the dataset provides data on the year built and year renovated, allowing for the examination of property age and recent updates as factors that could influence market value. Together, these variables are critical for understanding the drivers of house prices and provide a robust foundation for estimating the dependent variable, which is the price.

Fig. 1 Distribution of predictors

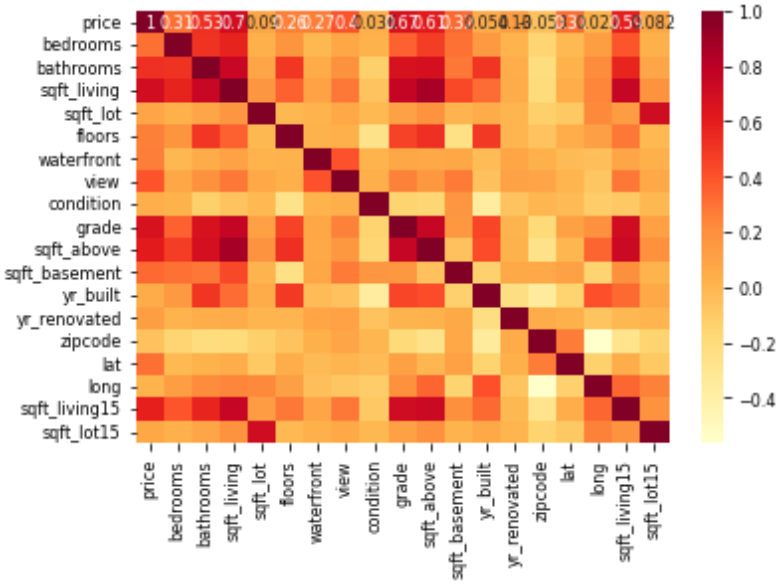


The Kings County dataset exhibits distinct distributional characteristics across its various features. A key observation is the pronounced right-skewness or positive skewness observed in the distributions of critical variables such as sale price, square footage, and lot size. This suggests that the majority of houses in the dataset cluster around the lower end of these measures, with a long tail extending towards higher values. In contrast, variables representing house condition and quality grades display more symmetric, normal-like distributions. This indicates a relatively balanced representation of houses across different conditions and grade levels, without a significant concentration at either the lower or higher end of the spectrum.

This pronounced right-skewness observed in the distributions of key numeric variables suggests that the dataset may present challenges for certain regression models. Linear regression, for instance, relies on the assumption of normally distributed residuals, which may be violated due to the skewed nature of the target variable and related features. In such cases, the model may underperform, particularly in predicting the prices of the high-value outliers that exist in the right tail of the distribution

The right-skewed nature of key numeric variables suggests a housing market dominated by relatively affordable properties, with a small number of high-priced outliers. The more symmetric, normal-like distributions observed for house condition and quality grade variables suggest that these features may be well-suited for linear modeling approaches. The balanced representation of houses across different condition and grade levels implies that these variables can be effectively utilized in linear regression models without introducing significant biases

Fig. 2 Correlation heatmap of predictors to price



The correlation heatmap illustrates the relationships between different features. The strongest correlation is observed between the price and the square footage of the living area, suggesting that larger homes tend to have higher prices. The heatmap also reveals a positive correlation between the number of bedrooms and bathrooms. Furthermore, there is a negative correlation between the year built and the price, indicating that older homes may be priced lower. The correlation values demonstrate the strength of each numeric feature's relationship with the price. For example, 'sqft_living' has the highest positive correlation, while 'sqft_lot' has a slight negative correlation. The heatmap shows that 'sqft_living' and 'grade' have the strongest positive correlations with 'price', while 'sqft_lot' exhibits a slight negative correlation. Given the correlation heatmap, features like 'sqft_living' and 'grade' display strong linear relationships with 'price', suggesting that a polynomial regression might effectively capture any nonlinear relationships.

Fig. 3 msno matrix to find missing values

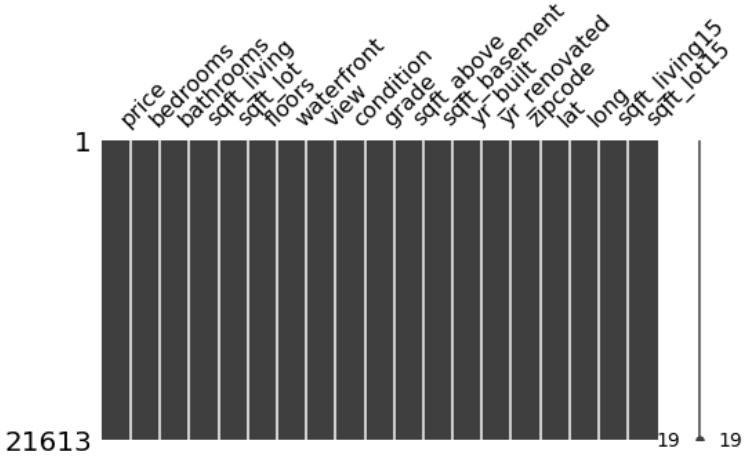


Figure 3 above shows there are no missing values in the dataset, which is crucial for ensuring that the relationship between variables produces consistent results. Not having missing data also ensures that all instances contribute fully to the training process to enhance the learning process which is essential for capturing underlying patterns accurately. Besides, missing data if not corrected can introduce bias into the models that are developed rendering the prediction ability of models inaccurate.

5. Results and Discussions

The exposition of the empirical results begins with examining the accuracy of the trained models. The algorithms utilized in this study include Multiple Linear Regression, Random Forest, Ridge Regression, Lasso Regression, and a Cubic Polynomial Regression. Each model was developed using Python and evaluated based on its Mean Squared Error performance metrics. MSE is a widely used and robust evaluation metric for regression models as it captures the average squared difference between the predicted values and the true values, providing a comprehensive measure of the model's performance (Botchkarev, 2019; Wang & Bovik, 2009). A lower MSE indicates that the model's predictions are closer to the true values, suggesting better overall fit and predictive accuracy. Table 2 shows the performance of our models as well as their predicted values. The difference between the MSE and the predicted values gives us insights into the magnitude of the error or deviation between the model's prediction and the true value. Thus, a smaller difference means that the model's predictions are closer to the true value, suggesting better model performance, and conversely, a larger difference indicates poorer model performance.

Table. 2

Estimated results for MLR, Lasso, Ridge, RF, Cubic polynomial

Model	Training MSE	Test MSE	Train MSE -Test MSE
Multiple Linear Regression	0.3113	0.2692	0.0421
Lasso	0.3113	0.2691	0.0422
Ridge	0.3113	0.2692	0.0421
Random forest	0.1364	0.0952	0.0412
Cubic Polynomial	0.1717	0.1297	0.0420

Prior to the development of the machine learning models, the dataset underwent a standardization process. This was done using the z-score standardization method, where each feature was transformed to have a mean of zero and a standard deviation of one. Standardization is crucial in machine learning as it ensures that each feature contributes equally to the distance calculations and model performance. This step is essential, particularly for algorithms sensitive to the scale of the input data, such as linear regression and regularized models (Ridge and Lasso), as it helps mitigate issues related to multicollinearity and improves convergence during optimization.

After standardizing the dataset, the data was split into training and test sets; 80% for training and 20% for testing. Thus, the training set consists of 17,290 samples, while the testing set has 4,323 samples. The rationale behind this split is to provide a clear distinction between the data used to train the model and the data reserved for testing its performance. In machine learning, it is imperative to evaluate the model on unseen data to gauge its generalization capabilities. By holding out a separate test set, we could ascertain how well a model performs in real-world scenarios, where predictions are made on data that was not part of the training process. This practice helps to prevent overfitting, where a model learns the training data too well, including its noise and outliers, resulting in poor performance on new data.

For hyperparameter tuning each of the five algorithms was trained on the training dataset, and 10-fold cross-validation(cv =10) was employed to ensure robust evaluation (Zhang & Yang, 2015). Cross-validation is a technique used to assess how the results of a statistical analysis will generalize to an independent dataset. In 10-fold cross-validation, the training dataset is divided into ten subsets. The model is trained on nine of these subsets and validated on the remaining one. This process was repeated ten times, with each subset serving as the validation set once. The average performance across all folds provides a more reliable estimate of the model's predictive capabilities. The choice of 10 folds strikes a balance between computational efficiency and the reliability of the validation process. Usually, fewer folds may lead to higher variance in performance estimates, while more folds can increase computational costs without significant gains in accuracy. Thus, by employing 10-fold cross-validation, we aimed to minimize bias and variance, ensuring that our performance metrics are robust and representative of the model's true capabilities.

Multiple Linear Regression (MLR) served as the foundational algorithm for the study, modeling the relationship between house prices and multiple independent variables. This model assumes a linear relationship between the predictors and the target variable, and we used the Ordinary Least Squares (OLS) method to estimate the model coefficients. The MSE for MLR was found to be 0.2692, which is relatively high and indicates that while the model captured some of the variance in the data, a significant amount of error remains in its predictions. This suggests that MLR, while simple and interpretable, struggles to model more complex relationships within the dataset, especially when issues like multicollinearity and others we discussed ealier in our description of the dataset are present.

The Random Forest algorithm, an ensemble learning method, constructs multiple decision trees during training and aggregates their predictions to improve generalization. In our study, Random Forest achieved the lowest MSE at 0.0952, outperforming all other models. This result highlights the effectiveness of ensemble methods in capturing complex patterns in the data. The algorithm’s ability to handle non-linear relationships and interactions between features contributed to its superior performance. Moreover, the feature importance metrics identified sqft_living: 0.3263, grade: 0.2659, lat: 0.1595, long: 0.0695, waterfront: 0.0348 as the most important variables that affect price. By limiting the tree depth to 15 and setting the seed to 0, the model strikes a

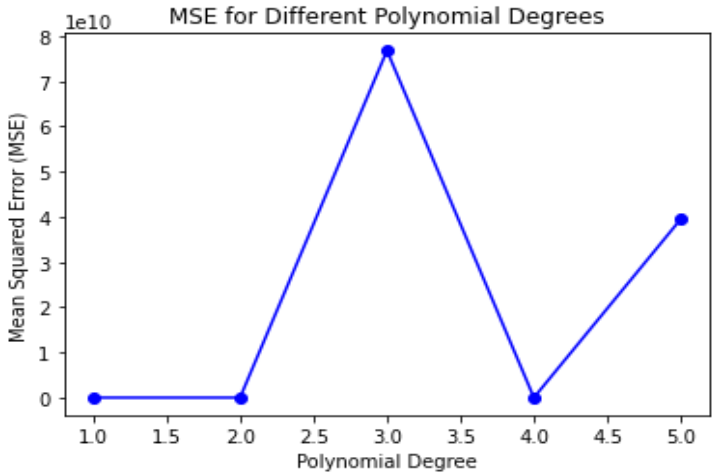
balance between bias and variance thereby improving the model's improved performance over its counterparts.

The MSE for Ridge Regression was also recorded at 0.2692, similar to that of the MLR model. This result suggests that while Ridge Regression helps to stabilize the estimates of coefficients, it does not significantly enhance the predictive power compared to MLR in this instance. The regularization parameter for the model was 50.0 which controlled the strength of the penalty, and was fine-tuned using cross-validation to optimize model performance. However, in this case, it appears that the underlying relationships in the dataset may not have benefitted from the regularization approach as much as anticipated.

Like Ridge, The MSE for Lasso Regression was also 0.2691, mirroring the results of MLR and Ridge Regression. This indicates that while Lasso Regression offers the advantage of variable selection, it did not provide a predictive edge in this particular analysis.This is evident from the fact that Lasso excluded only one variable(Sqft_ basement) out of nineteen. This means that the square foot of the basement did not contribute to house price prediction in this model. Because the dataset lacked an adequate number of irrelevant features that would have leveraged the inherent feature selection capabilities of Lasso regression, the performance of Lasso regression was found to be comparable to that of the more straightforward multiple linear regression (MLR) and Ridge regression models.

The MSE for the Cubic Polynomial Regression was found to be 0.1297, indicating a modest improvement over the linear models (MLR, Ridge, and Lasso) but still not as effective as the Random Forest model. The polynomial regression's capacity to capture nonlinear relationships is evident in its performance; however, it also raises concerns regarding overfitting, especially with a high degree polynomial. In Figure 4, it can be observed that the MSE decreases as the polynomial degree increases from 1 to 3, and then starts to increase again from degree 4 onwards. Hence, the lowest MSE is observed at degree 3, indicating that a 3rd-degree polynomial model provides the best fit for the house price dataset.

Fig. 4



After degree 3, the MSE starts increasing, suggesting that higher-degree polynomials (degree 4, 5 and 6) are likely overfitting the data and performing worse on the test set. Using this optimal polynomial for the prediction yields a test score of 0.1297.

Fig. 5 Comparison of Model Estimations

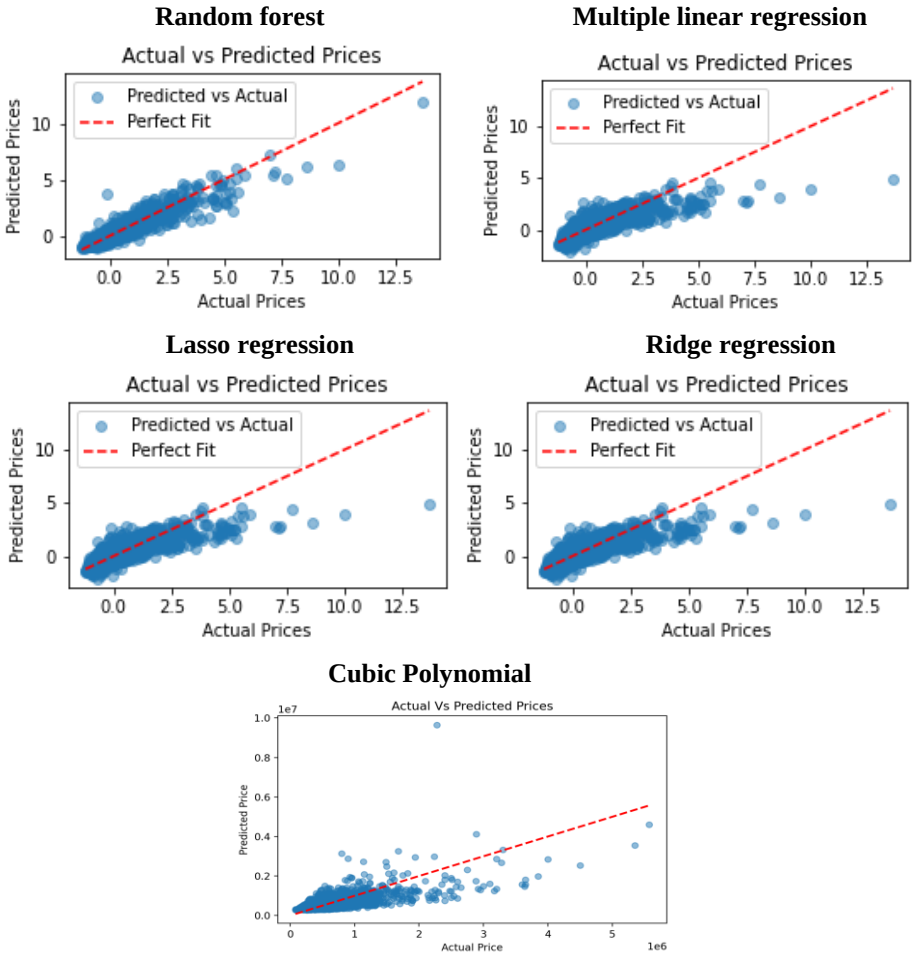


Figure 5 above shows a widening dispersion of prediction points with increasing prices indicates a decreased accuracy of the model for higher-priced homes. A significant concentration of predictions is observed in the lower price range, reflecting the predominance of such properties in the dataset. Furthermore, the model struggles to account for extreme values, particularly among very expensive homes, as evidenced by the heightened scatter observed at the upper end of the price spectrum. The Random Forest model emerged as the clear leader in terms of predictive accuracy, achieving an MSE of 0.0952. This performance highlights the strength of ensemble methods in managing complexity and capturing intricate relationships within the house price dataset which complexities include outliers, multicollinearity and others as observed and discussed in the data description section.

On the contrary, the linear models, while interpretable and straightforward, did not capture the complexities of the data as effectively as all exhibited identical MSEs of 0.2692, suggesting that the linear assumptions of these models may not have been suitable for the underlying data

structure. However, it can be observed from table 2 that all models show relatively small differences between their training and test MSE (around 0.042). This means that each model learned the underlying patterns in the training data and was able to generalize those patterns to new, unseen data with no significant overfitting issues. That said, Figure 4 shows most of the predicted values lie closer to the redline in the random forest model than the other models, indicating that random forest fits the study dataset than the other linear based models. Thus, our findings confirm that of a previous study by Ho et al. (2020) which concluded that Random forest algorithms give the best performance for predicting property prices. Thus although, the choice of machine learning algorithm is influenced by the size of the dataset, the computing power available and many other factors, ensemble methods, particularly Random forest, have proven to give more precise predictions of house prices where prediction accuracy is priority.

6. Conclusion

This study evaluated the predictive accuracy of several machine learning algorithms in estimating residential property prices in King County, Washington. Using a dataset of 21,613 property sales, we analyzed the performance of Multiple Linear Regression, Ridge Regression, Lasso Regression, Random Forest, and a Cubic Polynomial Regression model. Our results align with previous studies, confirming that ensemble methods, particularly Random Forest, yield superior predictive accuracy. With the lowest mean squared error (MSE) of 0.0952, Random Forest demonstrated the ability to capture complex, non-linear relationships between housing attributes and property prices more effectively than the linear-based models.

While Ridge and Lasso Regression contributed marginally to coefficient stability through regularization, they did not significantly enhance predictive power over Multiple Linear Regression in this dataset. Lasso's variable selection excluded only one feature, suggesting limited benefit from feature selection due to the lack of high-dimensional noise variables. The cubic polynomial model, optimized at degree three, highlighted the limitations of polynomial expansion, as degrees beyond three introduced overfitting and decreased model accuracy on the test set.

In line with prior research, the findings underscore the value of ensemble models for accurate property valuation, particularly when dataset variability and outliers are prominent. However, practical considerations such as model interpretability, computation time, and data complexity should guide algorithm choice. Random Forest, due to its robust handling of non-linearity and feature interactions, emerges as a promising tool for real estate appraisal, yet further research could explore hybrid models and advanced machine learning techniques for enhanced interpretability and precision.

References

Bastos, J. A., & Paquette, J. (2024). On the uncertainty of real estate price predictions. Journal of Property Research, 1–19. https://doi.org/10.1080/09599916.2024.2403998

Cawley, G., & Talbot, N. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J. Mach. Learn. Res., 11, 2079-2107. https://doi.org/10.5555/1756006.1859921.

Daly, J., Gronow, S., Jenkins, D., & Plimmer, F. (2003). Consumer behavior in the valuation of residential property. Property Management, 21, 295-314. https://doi.org/10.1108/02637470310508653.

Deppner, J., Von Ahlefeldt-Dehn, B., Beracha, E., & Schaefers, W. (2023). Boosting the accuracy of commercial Real estate appraisals: An Interpretable machine learning approach. The Journal of Real Estate Finance and Economics. https://doi.org/10.1007/s11146-023-09944-1

Harlfoxem. (2016). House sales in Kings County USA [Dataset]. Kaggle. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/data

Hjort, A., Pensar, J., Scheel, I., & Sommervoll, D. E. (2022). House price prediction with gradient boosted trees under different loss functions. Journal of Property Research, 39(4), 338–364. https://doi.org/10.1080/09599916.2022.2070525

Ho, W. K., Tang, B., & Wong, S. W. (2020). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Glauner, Petko Valtchev, Radu State arXiv (Cornell), 2018. Impact of Biases in Big Data Patrick https://doi.org/10.48550/arXiv.1803.00897

Matysiak, G., & Wang, P. (1995). Commercial property market prices and valuations: Analysing the correspondence. Journal of Property Research, 12, 181-202. https://doi.org/10.1080/09599919508724144.

Newell, G. (2024). Editorial: Through a referee's lens. Journal of Property Investment and Finance, 42(3), 221–222. https://doi.org/10.1108/jpif-04-2024-225

Wilson, I., Paris, S., Ware, J., & Jenkins, D. (2002). Residential property price time series forecasting with neural networks. Knowledge-Based Systems, 15(5–6), 335–341. https://doi.org/10.1016/s0950-7051(01)00169-1.

Tibshirani, R. J., & Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. The Annals of Applied Statistics, 3(2). https://doi.org/10.1214/08-aoas224.

Regression. (2005). In The MIT Press eBooks. https://doi.org/10.7551/mitpress/3206.003.0005

Lee, CF., Chen, HY., Lee, J. (2019). Multiple Linear Regression. In: Financial Econometrics, Mathematics and Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4939-9429-8_2

Lee, S. W. (2022). Regression analysis for continuous independent variables in medical research: statistical standard and guideline of Life Cycle Committee. Life Cycle, 2. https://doi.org/10.54724/lc.2022.e3

Hastie, T., Friedman, J., Tibshirani, R. (2001). Linear Methods for Regression. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-21606-5_3

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2010). Linear regression models. Wiley Series in Probability and Statistics, 9–76. https://doi.org/10.1002/9780470556986.ch2

Bayen, A. M., & Siauw, T. (2014). Least squares regression. In Elsevier eBooks (pp. 201–210). https://doi.org/10.1016/b978-0-12-420228-3.00013-0

Dionísio, A., Menezes, R., & Mendes, D. A. (2006). Entropy-Based Independence Test. Nonlinear Dynamics, 44(1–4), 351–357. https://doi.org/10.1007/s11071-006-2019-0

Wallach, D., & Goffinet, B. (1989). Mean squared error of prediction as a criterion for evaluating and comparing system models. Ecological Modelling, 44(3–4), 299–306. https://doi.org/10.1016/0304-3800(89)90035-5

Hoerl, A. E., & Kennard, R. W. (1981). Ridge Regression — 1980: Advances, Algorithms, and Applications. American Journal of Mathematical and Management Sciences, 1(1), 5–83. https://doi.org/10.1080/01966324.1981.10737061

Daoud, J. I. (2017). Multicollinearity and regression analysis. Journal of Physics Conference Series, 949, 012009. https://doi.org/10.1088/1742-6596/949/1/012009

Schreiber-Gregory, D. N. (2018). Ridge Regression and multicollinearity: An in-depth review. Model Assisted Statistics and Applications, 13(4), 359–365. https://doi.org/10.3233/mas-180446

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. Journal of Econometrics, 187(1), 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006

Hastie, T. (2020). Ridge regularization: an essential concept in data science. Technometrics, 62(4), 426–433. https://doi.org/10.1080/00401706.2020.1791959

García, J., Salmerón, R., García, C., & Del Mar López Martín, M. (2015). Standardization of variables and collinearity diagnostic in ridge regression. International Statistical Review, 84(2), 245–266. https://doi.org/10.1111/insr.12099

Emmert-Streib, F., & Dehmer, M. (2019). High-Dimensional LASSO-Based Computational regression models: regularization, shrinkage, and selection. Machine Learning and Knowledge Extraction, 1(1), 359–383. https://doi.org/10.3390/make1010021

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological), 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Lemhadri, I., Ruan, F., Abraham, L., & Tibshirani, R. (2019). LassoNet: A Neural Network with Feature Sparsity. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1907.12207

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. Journal of Chemical Information and Computer Sciences, 43(6), 1947–1958. https://doi.org/10.1021/ci034160g

Mentch, L., & Zhou, S. (2019). Randomization as Regularization: A Degrees of Freedom Explanation for Random Forest success. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1911.00190

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information Knowledge and Management, 14, 045–076. https://doi.org/10.28945/4184

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1811.12808

Pasandi, M. M., Hajabdollahi, M., Karimi, N., & Samavi, S. (2020). Modeling of pruning techniques for deep neural networks simplification. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2001.04062