House Price Prediction using Genetic algorithms and Tree-based methods for feature selection: The case of House Pricing in King County, USA

MSc Research Project
MSc in Data Analytics

Nihar Devidas Mhaske
Student ID: 20234813

School of Computing
National College of Ireland

Supervisor: Dr.Hicham Rifai

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Nihar Devidas Mhaske
Student ID:	20234813
Programme:	MSc in Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr.Hicham Rifai
Submission Due Date:	15/08/2022
Project Title:	House Price Prediction using Genetic algorithms and Tree-based methods for feature selection: The case of House Pricing in King County, USA
Word Count:	6947
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Nihar
Date:	17th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

House Price Prediction using Genetic algorithms and Tree-based methods for feature selection: The case of House Pricing in King County, USA

Nihar Devidas Mhaske
20234813

Abstract

The housing market is unstable and complex considering that it is impacted by the economic situation and a wide range of other factors. The housing market has the potential to contribute to economic stability. It is important to understand the housing market in to develop measures to promote sustainable and healthy housing values. The study aims to compare the performance of machine learning with two different feature selection techniques in predicting property sale prices. In this study, two machine models, Extreme gradient boosting and Random Forest, were implemented using a tree-based technique and genetic algorithms for feature selection. The model performed on the features selected by the tree-based approach has performed better than the genetic algorithm. In this study, the R2 score, mean absolute error, and cross-validation evaluation metrics were used to classify the model. Hyperparameter tuning was used to improve model performance. The implementation of models showed that the Random Forest model, which used a tree-based approach for feature selection, had the best performance in terms R2 score , MAE and RMSE compared to genetic feature selection models.

1 Introduction

The Americans are currently undergoing a housing crisis, with the shortage of 7 million property units for lower-class income residents. More than half a million people in the USA do not have a place to sleep. The property crisis is not the only main problem faced by Americans but affordability is a significant issue as well because many individuals cannot afford to purchase a home. As per the National Income Housing report in 2020, more than 70 % of low-income residents spend more than half of their income on housing (Islam; 2021). A housing market meta-analysis is required to comprehend the national problem. Property market analysis can provide significant information about the current situation of the national economy, general housing affordability, financial developments, and investments, in real estate market which can assist in the formulation of policies and strategies for building a cost-effective property market. Property price models can provide us with useful information about the factors that affect housing prices along with insights into future real estate market scenarios. It can, for example, determine which architectural and non-structural elements of houses contribute to higher sales. Housing pricing models can discover regional variations in housing prices. Real estate brokers also want a precise property price model to sell the property profitably and efficiently. Moreover, property price models will benefit in forecasting future property prices, helping the federal and government organizations to effectively regulate and control the entire market and support appropriate resource distribution before to any potential crisis. Buying a house is one of the biggest goals of an individual which requires a major amount of savings and net worth. Many studies reveal that overvalued houses are often sold in the USA, the main reason behind this is misleading advertisements and a lack of accurate and trust able information about the property market. Housing price modeling is one of the most critical parts of the examination of the housing market. Housing pricing models can give us important information about the factors influencing housing prices and estimates the upcoming market situations. In addition, it can identify the features of houses that increase the sales prices of the house.

The set of optimization algorithms termed genetic algorithms is based on the ideas of natural selection and evolution. The best feature combination can be identified by measuring every potential combination, and this becomes difficult as the number of available features increases over time. To eliminate unnecessary and irrelevant features, dimension reduction and feature selection approaches have been used(Schulte et al.; 2021). The Genetic algorithm has significantly improved machine learning techniques to predict changes in the real estate market and increase the accuracy of predictions of property prices. (Muneer et al.; 2022)

1.1 Motivation

The motivation behind this research has two primary reasons. Firstly, research help builder, buyers, and investors to know the valuation of property to make the best investment choices. It will help the seller by removing the auction expenses and helping buyers to know house prices estimated by the model. Secondly, genetic algorithms and tree-based methods used for feature selection will help to improve the performance of the models. Best features selected by tree-based method and genetic algorithms will help builders to know important features and which features are most valued by the customers to increase profit. A novel optimization technique would improve upon existing research in the fields of property valuation.

1.2 Research Question

How well a machine learning model perform using genetic algorithm for feature selection compared to tree based method?

1.3 Objective

Objectives 1 review all the related work done in this domain followed by objectives 2 and 3 focus on pre-processing the data and performing exploratory analysis. Objective 4 select the best feature for the dataset to predict house price using genetic algorithms and tree-based methods. Objectives 5 and 6 comprise machine learning, data mining, genetic algorithms, tree-based techniques, and comparison between machine learning models. The major investigation of this research is using novel genetic algorithms and tree-based method to identify the most accurate machine learning model to predict house prices. By performing exploratory analysis major insights of the features were analyzed.

- **Objective 1:** Investigate literature review of property price prediction using genetic algorithm and tree based method for feature selection
- **Objective 2:** Data cleaning, dealing with missing values and feature engineering.
- **Objective 3:** Exploratory data analysis of price prediction and transformation.
- **Objective 4:** Implementation of feature selection technique using genetic algorithms and tree-based techniques
- **Objective 5:** Implementation of machine learning models
- **Objective 6:** Comparison of the model and interpret findings .

1.4 Structure of the paper

In this paper, we will have a better understanding of various methods and their findings. Each section in this research will help to learn more about that section. In Section 2, different research paper on this domain was studied which gave insight into the techniques used in the house price prediction domain. Section 3 discusses the methodology proposed in this paper. Section 4 Design Specification and the process flow of the research is discussed. Section 5 Implementation of feature selection technique and machine learning model is discussed. Section 6 includes Evaluation and Discussions of the results. Section 7 includes conclusion of the research and future work

2 Literature Review

In this literature review, feature selection techniques used to optimize machine learning and house value prediction models are analysed and evaluated. This research focus on genetic algorithm and tree-based method for feature selection, so the performance of GA algorithms is also studied. The analysis of the property valuation domain and a review of effective optimization techniques and machine learning are discussed in this section.

2.1 Feature Selection Techniques

The feature selection technique selects the best feature for modeling, analyzing, and building a machine learning model. The objective is to reduce the number of parameters in the model while retaining the most predictive feature and removing the lower predictive feature. Feature selection is essential as most models do not deal with a lot of irrelevant features, which can increase model noise and may lead to overfitting (Islam; 2021). As fewer features don't even adequately describe the data, generalizing a model with fewer data can be difficult. Using feature selection methods to identify the key factors that are more or less significant in predicting property values(Ahtesham et al.; 2020). The accuracy of valuation is directly impacted by the selection and interpretation of feature variables. (Niu and Niu; 2019)

The three different feature selection methods are wrapper methods, filter methods, and hybrid methods. During the pre-processing phase, filtering techniques are frequently employed. Filter methods primarily use fundamental statistics including. Information gain, correlation, chi-square test, variance, and correlation coefficient. By using the correlation matrix evaluation of dependent and independent features was performed on the calculated correlation matrix. This study identify which variables were crucial for predicting house prices.(Abdul-Rahman et al.; 2021). Using wrapper techniques adds or delete predictors, wrapper approaches compare many models to determine the best combination that maximizes model performance. The ANOVA, Backward, Forward, and stepwise selection these techniques are used in wrapper methods. The stepwise selection approach integrates the forward and backward selection techniques. It begins without any predictors and adds the most predictor variables one at a time, by the sequential order of importance. (Phan; 2018). The hybrid selection approach has built-in methods for selecting features that are integrated into the process of learning using embedded methods. Hybrid selection techniques combine two different procedures. Tree Based technique uses boosting techniques and random forest feature selection methods. L1 and L2 methods are known as regularization feature selection methods. The top 10 significant variables were evaluated by L1 regularization (Yan and Zong; 2020) and a random forest tree-based approach were used for feature selection using MAPE(Hong et al.; 2020)

Analysis of previous studies has revealed that feature selection methods are one of the essential phases in building machine learning models. The filter approach has the advantage of being extremely quick and affordable to compute, being effective at eliminating redundant, correlated, and duplicate features. The drawback of filter methods is that they do not eliminate multicollinearity. The advantage of wrapper methods over the filter technique is that they offer the best set of variables to train the model, which results in high accuracy than the filter technique while being more computationally expensive. A drawback of wrapper and filter has been overcome by hybrid methods. These methods are comparable to filter methods in that they take into account several features and are quicker and more accurate than filter methods.

2.2 Feature Selection using Tree-Based Methods

The decision tree is the process of feature selection which is a tree-based feature selection approach. When selecting the features of split sampling of each layer, each feature is computed following specific specifications, and the feature which important is selected. The decision tree method's main advantages are high accuracy and stability. The downside is that it is sensitive to overfitting, the overfitting problem can be solved by performing a cross-validation technique (Zhou et al.; 2021). A Gini Impurity-based weighted Random Forest model was used in this research to identify the most important features. The feature importance rankings were computed by varying their weight within the Random Forest algorithm for an imbalanced class researcher used the Gini impurity to split the trees. The Adaptive Boosting, Gradient Boosting Tree, Long Short Term Memory, Decision Tree, and Gated Recurrent Unit model were performed on two datasets. (Disha and Waheed; 2022). XGBoost is an ensemble learning technique that combines many weak classifications. By creating numerous CARTs parallel, this method can boost computing speed. Moreover, the technique does not require feature standardization. A sparse perception method used by XGBoost will learn its split pattern for samples having missing data. But, prior to split, XGBoost needs to evaluate the importance of the features, which increases the accuracy of the model(Li et al.; 2022). For the feature selection phase, an Gradient Boosted Regression Trees (GBRT) is used by the reseachers (Ramos-González et al.; 2017). The GBRT algorithm is more commonly employed for classifications than the regression problem. The proposed architecture in this research, Gradient Boosting is selected as a feature selection approach. This approach reduces the number of genes and the best feature selected by GBRT have significantly increased the performance

2.3 Feature Selection using Genetic Algorithms

A method for solving optimization issues using natural selection is a genetic algorithm. Selection, Crossover, Population Initialization, Fitness evaluation, and Mutation are the five main components of Genetic algorithms. For the most effective selection of the feature set from a multi-character collection, an improved genetic algorithm was suggested by the researcher. With this method, the chromosome is separated into different classifications. Then, different crossover and mutation operators are used to determine the categories to remove any invalid chromosomes. For medical de - noising, CNN methods were implemented by initializing the population of hype-parameter optimization of genetic algorithms parameters(Dong et al.; 2018). Meta-heuristic search methods, including Particle Swarm Optimization and Genetic algorithm, have been selected for feature selection because an exhaustive search would be computationally extremely costly. A few iterations of the genetic algorithms and support vector machine methods have been used for feature selection in various research domain areas. Genetic algorithms and support vector machine method was investigated for the categorization of hyperspectral pictures(Jadhav et al.; 2018). For feature selection, a genetic algorithm was used in this research. The outliers from the dataset were eliminated using the k means clustering algorithm, and the best features were then selected using a Genetic Algorithm. The classifier models of the Naive Bayes, Support Vector Machine, k-nearest neighbor, and, decision tree, were implemented using these selected features by genetic algorithms(Anirudha et al.; 2014). The proposed technique modifies the initial population to select the best features, then applies an information-based ranking of variables to select the best feature set. Second, by modifying the population, mutation, and crossover of genetic algorithms parameters, this measurement directs the evolution of GA.

2.4 Related Work on King County, USA Dataset

The researcher used different regression techniques, including Lasso, Ridge, Multiple regression, Elastic Net, and Ada Boosting model, to predict the price of houses sold in King County, USA. The model's performance was measured using its accuracy, and Mean Square Error(Madhuri et al.; 2019). To predict the sale price of houses in King County, authors (Azimlu et al.; 2021) search for solutions that mixture of supervised ML methods and clustering algorithms. Clustering of data was performed in the initial stage using clustering methods like k-Means and DBSCAN clustering. After clustering, machine learning models were used for each cluster such as Support vector regression, Linear regression, and ANN. The researcher (Islam; 2021) used the King County dataset to predict house sale price study. Stepwise selection, a filter method, was applied by the researchers as a feature selection approach. In the data processing processes, a researcher discovered that the price feature is not distributed normally then log transformation is performed on price features to make it normally distributed. Many algorithms were implemented such as Support Vector Regression, Geographically Weighted Regression, and XgBoost model to evaluate the performance of the model Mean Square Error, Root mean square error, and R -squared were used.

2.5 Conclusion and Identified Gaps

The related work determined that the most successful algorithms for predicting house prices in King County were ensemble machine learning techniques as shown in Table.1. The performance of the model was increased using the optimization technique for feature selection. Many studies across this domain have implemented embedded, filter, or wrapper methods for feature selection. This research has novel genetic algorithms methodologies and state of the art across work done on the King County dataset, no study has implemented genetic algorithms on this dataset.

Table 1: Comparison of Techniques Used

Citation	Dataset	Machine Learning Models	Feature Selection Techniques
(Madhuri et al.; 2019)	King County, USA	Multiple Linear Regression, Ridge Regression, LASSO Regression, Elastic Net Regression, Ada-Boosting Regression, Gradient Boosting.	Tree-Based Feature Selection
(Azimlu et al.; 2021)	King County, USA	Linear Regression, Polynomial Model, Support Vector Regression, Artificial Neural Networks, Symbolic Regression	Regularization Feature Selection
(Islam; 2021)	King County, USA	Geographically Weighted Regression, Random-Effect-Eigenvector Spatial Filtering, RFESF-based SVC, Support Vector Regression, XGBoost ,Support Vector Regression Model	Stepwise Feature Selection
This Research	King County, USA	XGBoost, Random Forest	Genetic Algorithms and Tree-Based Methods for Feature Selection

3 Methodology

For this research, a Knowledge Discovery Database (KDD) methodology has been used. The KDD methodology follow steps shown in Fig.1 to get knowledge for the datasets. The preliminary methodological procedures will be briefly outlined in this section.

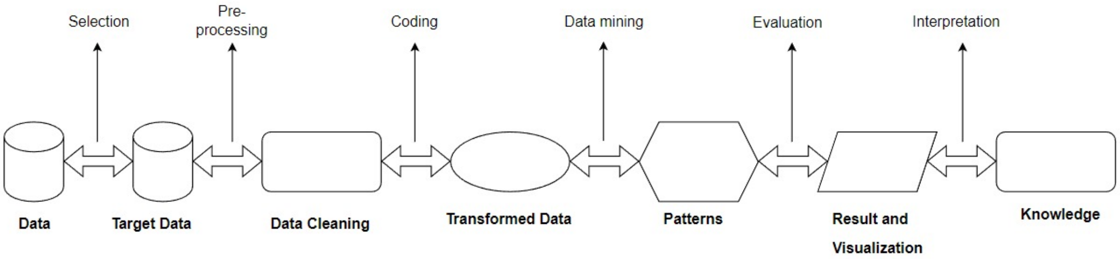

Figure 1: KDD Methodology

3.1 Data Selection

Data Selection- The data used in this research is obtained from the Kaggle website. The King County house sale prices are featured in this dataset. Properties sold between May 2014 and May 2015 have been included. Dataset consists of 21 features and 21,613 rows.

3.2 Data Pre-processing and Transformation

Using the pandas library, the dataset was imported. The dataset didn't have any missing values. After analyzing the data's statistics, we discovered that the target variables are not normally disturbed, after removing all the outliers of the price variable was normally disturbed. For further analysis, date features were transformed to year, month, and day. Each feature performed exploratory data analysis to acquire insight into the data, and outliers were removed using a boxplot. To identify the expensive location, longitude and latitude features were plotted using a scatterplot. Unwanted columns and highly correlated features were removed using a correlation matrix. To normalize the data for the stage of building the model, Standard Scaler was used. After this process, genetic algorithms and tree-based methods were used to select the best features for analysis.

3.3 Data Modeling

3.3.1 Extreme Gradient Boosting

Cehn and Guestrin (2016) created extreme gradient boosting as an ensemble of Classification and Regression Tree techniques. Extreme gradient boosting is a more advanced variation of gradient boosting. The key strengths of this technique are its speed and performance in comparison to other methods for machine learning. It combines the concepts of parallel computing, cache enhancement, which is accomplished by storing all of its statistics in memory, and out-of-storage computation, which maximizes the memory to allow it to perform on data that is more than the amount of the memory. The XGBoost model performs more quickly because of all these features. It includes regularization, which assists in preventing the model from overfitting. It has capable of handling the missing value.

3.3.2 Random Forest

Random forest is a method implemented by Breiman and Adele Cutler that combines the results of many decision trees to get a single outcome. The number of tree increases, and high variance and overfitting problems is prevented by random forests. The random forest can accurately perform both classification and regression problems. In this approach, bags are generated using a set of columns and rows, then each bag has a decision tree assigned to it.

3.4 Evaluation

The evaluation metrics (Cross-validation, Mean Absolute error, R2 score and Root Mean Square Error) were used after the machine learning models were trained and tested on the data to evaluate the model's performance and validate a hypothesis. In Section 6, the evaluation metrics are described in more detail.

4 Design Specification

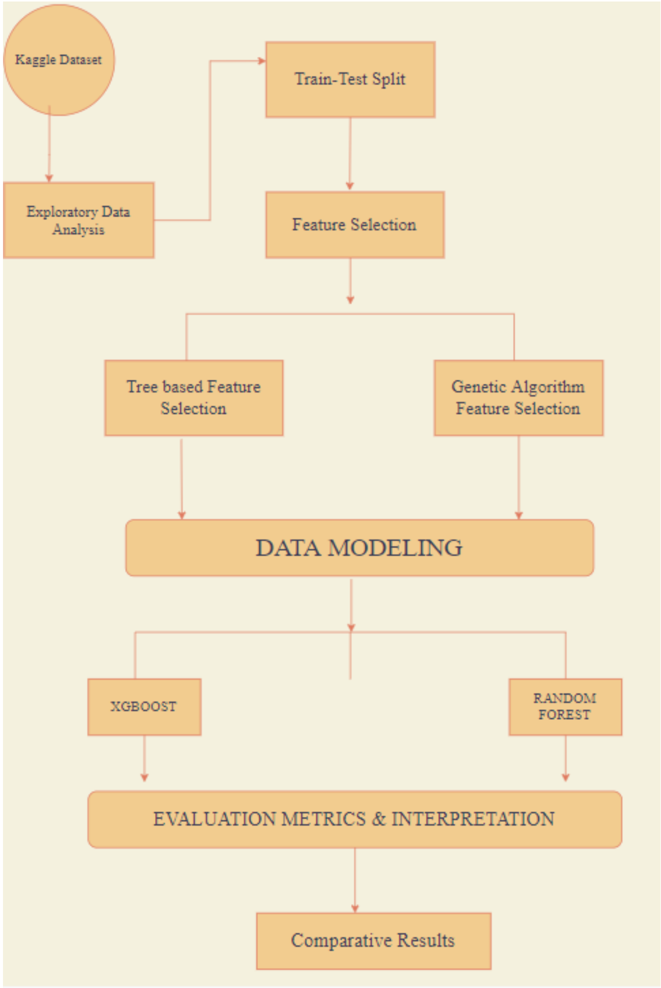
Kaggle Dataset
Train-Test Split
Exploratory Data Analysis
Feature Selection
Tree based Feature Selection
Genetic Algorithm Feature Selection
DATA MODELING
XGBOOST
RANDOM FOREST
EVALUATION METRICS & INTERPRETATION
Comparative Results

Figure 2: Process Flow Diagram

The steps used to address the research question are described in the process flow diagram in Fig.2. Due to the simplicity of use and the ease of access of several libraries like pandas, NumPy, Sklearn, seaborn, and Matplotlib were used for this analysis, Python programming was the fundamental programming tool used for the development. The Jupyter Notebook (IDE), was used to perform data transformation, pre-processing, and

Exploratory Data Analysis steps. After performing all the pre-processing steps data was standardized and then spitted into train and test sets using the sklearn library. The best features were then selected using feature selection techniques such as tree-based methods and genetic algorithms. Machine learning models like random forest and XgBoost were implemented after selecting the best feature for the data. Seaborn and Matplotlib, two Python libraries, were used to show the findings and results.

5 Implementation

The implementation of the research is covered in this section. This section describes all techniques used to accurately estimate house price sales. Two feature selection methods have been used, and they have been compared. The accuracy of the machine learning model has been improved by the use of hyperparameter tuning.

5.1 Feature Selection Techniques

The best feature is selected using a feature selection technique for the building model. This study's main goal is to select features using genetic algorithms and tree-based methods. The comparison of feature selection techniques used in the study is evaluated to know which feature selection technique has performed better in this research.

5.1.1 Tree based feature selection

Tree-based feature selection is classified as an embedded method. Tree-based feature selection techniques that are widely used are Random Forest, Extra Tree, Decision Tree, and xgboost. In this study, random forest and XGBoost were used to determine the important features.

Random Forest for Feature Selection: The random forest method is enhanced through bagging, and it applies the CART decision tree to a weak learner. Since the efficiency of the Random Forest algorithm, it also performs well in feature selection and measuring feature importance. The importance score assigned to a feature by Random Forest reflects its role to build a decision tree. The more frequently a feature is used as a dividing attribute across all trees, the more significant the feature contributes. If any noisy data gets added to the features the precision of the bags reduces which indicates the features have a strong impact on the results (Liu and Song; 2021).

Extreme Gradient Boosting for Feature Selection: XGBoost evaluates feature importance by measuring and categorizing each feature in the dataset. The relevance of features in a single decision tree is measured by the amount of high performance measured at each feature split point, and also the nodes parameter is for weighting and recording the timings. The Gini purity of an identified split node is used to evaluate performance. Then, the importance score is calculated by weighting the outcomes of each attribute across all boosted trees.

5.1.2 Genetic Algorithm for Feature Selection

A method for natural selection-based optimization is a genetic algorithm. Implementation of a genetic algorithm consists of four main steps shown in Fig.3 which are selection, crossover, mutation, and generation. For feature selection, the first step is to generate a population based on subsets of the possible features. The genetic algorithm is one of the stochastic feature selection methods based on evolution and natural selection. The process of natural evolution serves as an inspiration for a genetic algorithm, which is a heuristic optimization method. Implementation of a genetic algorithm consists of four main steps, selection, crossover, mutation, and generation. The first stage is to initialize the population's individuals. The population is created by the number of features available in the data in our case 13. After the population is initialized, each individual in the population has been assigned the fitness value. The fitness value selects the best score by training one model which we have used is random forest and the best population with an important score was collected.

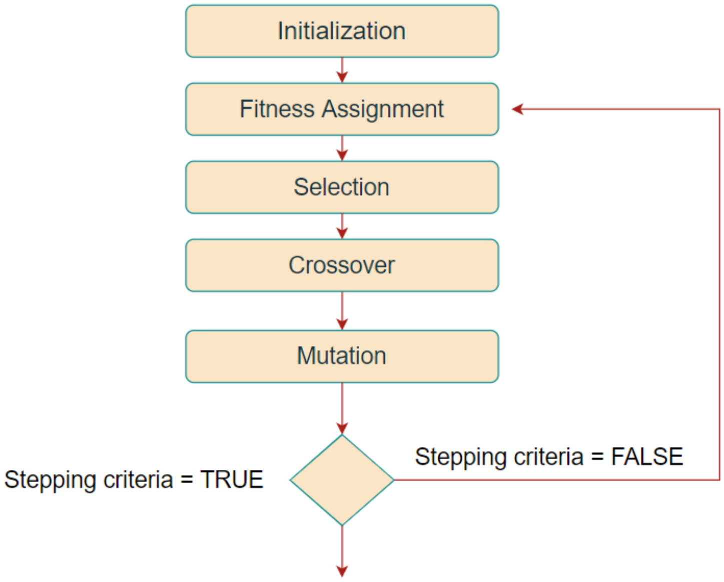

Figure 3: Genetic Algorithm Steps

Selection: After completing a fitness, the selection operator selects the individuals to recombine for the next generation. In the selection process, we select the preferred size of the population. It expects two parameter values. One is an output of fitness core function another one size of the number of best score giving population.

Crossover: The cross-over process increases the selected population size. Each time order wise it takes two populations. In that two populations, one is considered as child 1 another one is child 2. It finds half the population size. In our case 6. then it does split and merge the first 6 elements from child 1, in child 2 after elements of 6 and it creates new population.

Mutation: This mutation function is used to mutate the elements in the populations. In this process, the function generates two random numbers between feature size 13. After this, it takes one population and two random numbers if the position of the number taken in the population if true converts into false and if false converts into true.

Generation: Generation function used to collect the best 5 populations in that given 5 generations with importance score. The generation function selects the population with the best score from the five generations.

5.2 Hyperparameter

The most time taking procedure of our analysis was model optimization. Hyperparameters in machine learning are the parameters of ML techniques that achieve the best results when evaluated on a testing set and they are assigned before training. The best parameter for random forest and XGBoost is selected by hyperparameter tuning. Hyperparameter optimization selects a set of hyperparameters that results in an efficient model that minimizes a prediction error and improves the accuracy of the model. It is important to know how to optimize them to obtain maximum performance. Hyperparameter used in model building steps is shown in Table.2.

Table 2: Hyperparameter Used for Model Building

Hyperparameters	Model	Breif
Learning rate	XGBoost	New trees is generated using previous trees series, so that residual error is corrected
min child weight	XGBoost	If a trees partition process returns an leaf node with a total of instance weights less than min child weight, the building cycle will stop partition further.
gamma	XGBoost	This only allows splitting when the loss function has a positive reduction. It indicates the minimum loss reduction required to perform a split.
max depth	Random forest and XGBoost	Its avoid the problem of overfitting
n estimators	Random forest and XGBoost	The number of decision or stumps is derived

6 Evaluation

A comprehensive analysis was performed on all four models after selecting the important feature using feature selection techniques discussed in the section of implementation. To improve accuracy, hyperparameter tuning was performed. The data was divided into 75-25 for train and testing. The models were evaluated using the R2 score, mean absolute error, Root Mean Square error and cross-validation score.

R2 Score: R2 score is know as R-squared which compares model predictions to the mean of the targets. The value range between negative infinity to 1.

Mean Absolute Error: MAE is the average of the absolute difference between predictions and actual values. It gives an idea how wrong the model prediction are made.

Root Mean Square Error: The standard deviation of the residuals is the Root Mean Square Error (RMSE). Residuals measure how far data points are off the regression line and how far the residuals are spread out.

Cross Validation: Cross- Validation is simply a re sampling method that ensures our model's accuracy and efficiency on unseen data. The purpose of cross-validation will be to evaluate the model's capability to predict data that was not utilized in its prediction, avoid problems like over fitting or selection bias, and provide insight into how the model will apply to an dataset.

When evaluating the model performance on the data, calculating both the R2 score and RMSE is helpful since each metric gives different insights. The main reason to use RMSE is that it indicates the average distance between the regression model's actual value and the predicted value. In this research, R2 is used to inform how effectively the predictor factors explain the variance in the predictor variables. The cross-validation score is used to prevent overfitting problem. These metrics for the four models prediction were employed in this study to evaluate the difference between the predicted and actual values. The model's performance increases as the value of the performance metrics like MAE, and RMSE decrease. The prediction accuracy is represented by R2.

6.1 Extreme Gradient Boosting

6.1.1 Extreme Gradient Boosting Model using Features selected by Tree-based method

The randomized search was used to find the best parameter to build the model. The parameter selected after hyperparameter tuning is shown in Table.3. The XGBoost model achieved an R2 score of 0.842 % on test data, 0.82% on cross-validation data, a mean absolute error value of 8.22 and root mean square error of 13.99. The Fig.4 shows the graph of actual value of the price and the price predicted by the model. The value predicted by the model is accurate compared with actual value. The important features selected by the XGBoost model to build the model were grade, waterfront, sqft_living, lat, view, yr_built, sqft_living15, and price.

6.1.2 Extreme Gradient Boosting Model using Features selected by Genetic Algorithms

The XGBoost model was applied to the features selected by Genetic algorithms. To improve the model's performance, hyperparameter tuning was performed using a ran-

Table 3: Parameter Used for XGBoost Model Using Tree-Based Method

Parameters	Value
base_score	0.5
gamma	0.4
learning_rate	0.30
max_depth	8
min_child	1
n_estimators	100

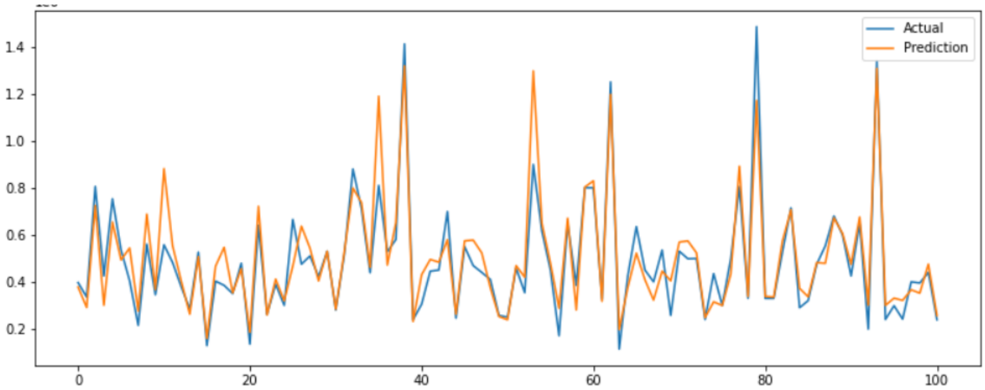

Figure 4: XGBoost Model Accuracy Using Tree-Based Method

domized search and the selected parameter is shown in Table.4. The model achieved an R2 score of 0.841, a cross-validation score of 0.81, an MAE value of 8.33 and RMSE of 14.04 . The Fig.5 depicts the actual value and predicted value of the model and there is no significant difference between actual and predicted value. The XGBoost model using feature selected by tree-based method perform slightly better than genetic algorithms.

Table 4: Parameter Used for XGBoost Model Using Genetic Algorithm

Parameters	Value
base_score	0.5
gamma	0.3
learning_rate	0.20
max_depth	6
min_child	1
n_estimators	100

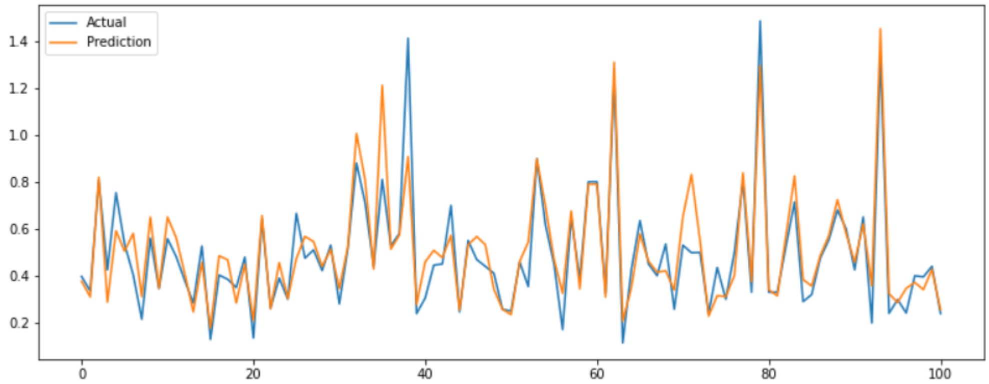

Figure 5: XGBoost Model Accuracy using Genetic Algorithm

6.2 Random Forest

6.2.1 Random Forest Model using Features selected by Tree-based method

Randomized cross-validation is used for hyperparameter tuning to identify the parameter that will increase the model's accuracy. Table.5 shows the best parameter selected by random search. The random model has an R2 of 0.843 % a cross-validation score of 0.82% a mean absolute error of 8.08 and RMSE of 13.96. The Fig.6 shows the prediction difference between the actual and predicted values. The graph show that the predicted values are accurate predicted. The random forest model used grade, sqft living, lat, yr built, sqft living15, waterfront, sqft lot, and price as important features to predict house sale price.

Table 5: Parameter Used for Random Forest Model Using Tree-Based Method

Parameters	Value
max_depth	20
min_sample_leaf	1
min_sample_split	2
n_estimators	60

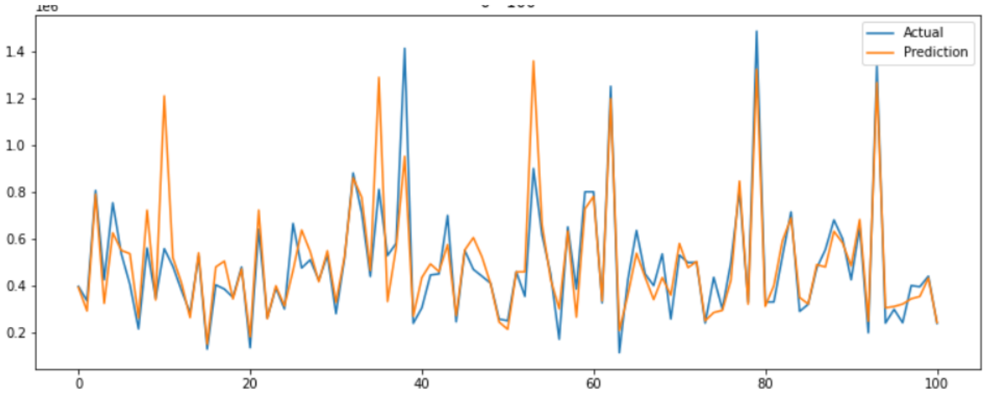

Figure 6: Random Forest Model Accuracy Using Tree-Based Method

6.2.2 Random Forest Model using Features selected by Genetic Algorithms

The Random Forest model was built on the features selected by Genetic algorithms. To improve the model's performance, hyperparameter tuning was performed using a randomized search. The random search cv was performed to select the best parameters to achieve better results. The parameter used for model building is shown in Table.6. The Random Forest model achieved an R2 score of 0.8394% a cross-validation score of 0.80% a mean absolute error of 8.55 and RMSE OF 14.04. The Fig.7 shows the difference between the actual value and value predicted by Random Forest model. The important features selected by genetic algorithms to predict house sale price were bedrooms', 'sqft_lot', 'floors', 'waterfront', 'grade', 'sqft_basement', 'yr_built', 'yr_renovated', 'lat', 'sqft_living15', 'sqft_lot15', 'price'. Random Forest using tree-based method perform better compared with RF model using genetic algorithms features selection.

Table 6: Parameter Used for Random Forest Model Using Genetic Algorithm

Parameters	Value
max_depth	20
min_sample_leaf	1
min_sample_split	2
n_estimators	100

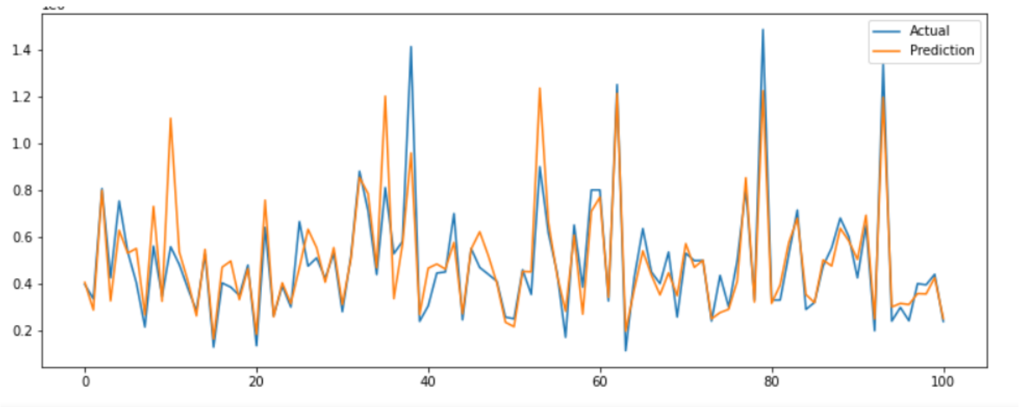

Figure 7: Random Model Accuracy Using Genetic Algorithm

6.3 Discussion

The main objective of this research was to compare two feature selection techniques and predict house sale prices based on model evaluation using r2 scores, mean absolute error, root mean sqaure error, and cross-validation. This study implemented two machine learning models on two different types of the feature selection approach. Table.7 and Table .8 shows the evaluation of the models perform on the feature selection techniques. Three sets of the importance of feature combinations were selected using the tree-based feature selection techniques with XGBoost and Random Forest model, while another set was selected using genetic algorithms. The Random Forest model built in this study using the Tree-Based technique predicted the residual value with the best accuracy of 84.3% followed by the accuracy of the XGBoost model using the Tree-Based method is 84.2%Ṫhe MSE and MAE result back up the accuracy results. The Random Forest Tree-based model has a lower MAE of 8.08% whereas the XGBoost tree-based method has an MAE of 8.22%İn this study, the Random Forest model outperformed all other models using a tree-based method. In this work, the comparison of two feature selection techniques revealed that the tree-based method outperformed genetic algorithms with small margin . When comparing two genetic algorithm models, XGBoost performed better than random forest with less MAE result as shown in Table.8.

Table 7: Evaluation Metrics - Feature Selected by Tree-Based Methods

Model	R2Score	Validation Score	Mean Absolute Error	Root Mean Square Error
Random Forest	0.843	0.82	8.08	13.96
XGBoost	0.842	0.82	8.22	13.99

Table 8: Evaluation Metrics - Feature Selected by Genetic Algorithm

Model	R2Score	Validation Score	Mean Absolute Error	Root Mean Square Error
Random Forest	0.839	0.80	8.55	14.04
XGBoost	0.841	0.81	8.33	14.12

7 Conclusion and Future Work

This investigation found which features are important for predicting the house prices. The novelty of this research was to implement genetic algorithms for feature selection. This research has consequences for optimizing tasks that require high computational components and the results of genetic algorithms can be better if performed on a higher specification system as it needs lots of memory allocation which is difficult to execute on a local system. To solve this gap, many researchers have used ensemble machine learning algorithms with limited hyperparameter tuning and dealt with overfitting problems in past work. Random Forest bagging techniques and XGBoost boosting techniques are used both have high compute power and perform well with noisy data. For feature selection, many researchers have employed wrapper, filter, and embedding techniques in previous work. Genetic feature selection is applied in this study to improve model accuracy. The model's results using a genetic algorithm and a Tree-based technique have performed better compared to related work done on this dataset.

Even though the investigation focuses on house sale price prediction, the data used for the study was limited to King County USA. As a result, the findings of this study may not apply to regions with different property market dynamics. The performance of various machine learning models was assessed in this study, even though only one dataset was used. To gain a better understanding of the unique method utilized in this study, it needs to be tested on multiple datasets. Genetic feature selection models that could have surpassed tree-based methods were not examined with the best parameter because of the computational expense. The number of a generation that genetic algorithms could have run was restricted by computational power. This has impacted the result of genetic algorithms. Based on this research it can be concluded that the systems with low computational power or iot based devices, which is the core for real time systems should make a use of machine learning algorithms that are computationally less intensive and can be executed on any resources constrained systems without any lag

8 Acknowledgement

I would like to thank my supervisor, Dr.Hicham Rifai, for his guidance, consistent support, and giving feedback throughout the research, and especially for suggesting me to work on genetic algorithms for feature selection to add novelty to my research. I would also want to thank Dr.Anu Sahni for the notes and resources she shared during the course, which were incredibly helpful in conducting this research. I am grateful for the opportunity to work on this project and contribute to the field of data science. I also want to thank my family and friends for always supporting me and pushing me to achieve my goals.

References

Abdul-Rahman, S., Zulkifley, N. H., Ismail, I. and Mutalib, S. (2021). Advanced machine learning algorithms for house price prediction: Case study in kuala lumpur, *International Journal of Advanced Computer Science and Applications* **12**(12).

Ahtesham, M., Bawany, N. Z. and Fatima, K. (2020). House price prediction using machine learning algorithm-the case of karachi city, pakistan, *2020 21st International Arab Conference on Information Technology (ACIT)*, IEEE, pp. 1–5.

Anirudha, R., Kannan, R. and Patil, N. (2014). Genetic algorithm based wrapper feature selection on hybrid prediction model for analysis of high dimensional data, *2014 9th International Conference on Industrial and Information Systems (ICIIS)*, IEEE, pp. 1–6.

Azimlu, F., Rahnamayan, S. and Makrehchi, M. (2021). House price prediction using clustering and genetic programming along with conducting a comparative study, *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 1809–1816.

Disha, R. A. and Waheed, S. (2022). Performance analysis of machine learning models for intrusion detection system using gini impurity-based weighted random forest (giwrf) feature selection technique, *Cybersecurity* **5**(1): 1–22.

Dong, H., Li, T., Ding, R. and Sun, J. (2018). A novel hybrid genetic algorithm with granular information for feature selection and optimization, *Applied Soft Computing* **65**: 33–46.

Hong, J., Choi, H. and Kim, W.-s. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in south korea, *International Journal of Strategic Property Management* **24**(3): 140–152.

Islam, D. (2021). *Housing Price Modeling: An Eigenvector Spatial Filtering Based Machine Learning Approach*, Central Michigan University.

Jadhav, S., He, H. and Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating, *Applied Soft Computing* **69**: 541–553.

Li, H., Shi, L., Gao, W., Zhang, Z., Zhang, L., Zhao, Y. and Wang, G. (2022). dpromoter-xgboost: Detecting promoters and strength by combining multiple descriptors and feature selection using xgboost, *Methods* .

Liu, Z. and Song, J. (2021). Comparison of tree-based feature selection algorithms on biological omics dataset, *2021 The 5th International Conference on Advances in Artificial Intelligence (ICAAI)*, pp. 165–169.

Madhuri, C. R., Anuradha, G. and Pujitha, M. V. (2019). House price prediction using regression techniques: a comparative study, *2019 International conference on smart structures and systems (ICSSS)*, IEEE, pp. 1–5.

Muneer, S. M., Alvi, M. B. and Rasool, M. A. (2022). Genetic algorithm based intelligent system for estate value estimation, *International Journal of Computational and Innovative Sciences* **1**(1).

Niu, J. and Niu, P. (2019). An intelligent automatic valuation system for real estate based on machine learning, *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, pp. 1–6.

Phan, T. D. (2018). Housing price prediction using machine learning algorithms: The case of melbourne city, australia, *2018 International conference on machine learning and data engineering (iCMLDE)*, IEEE, pp. 35–42.

Ramos-González, J., López-Sánchez, D., Castellanos-Garzón, J. A., de Paz, J. F. and Corchado, J. M. (2017). A cbr framework with gradient boosting based feature selection for lung cancer subtype classification, *Computers in biology and medicine* **86**: 98–106.

Schulte, R. V., Prinsen, E. C., Hermens, H. J. and Buurke, J. H. (2021). Genetic algorithm for feature selection in lower limb pattern recognition, *Frontiers in Robotics and AI* p. 324.

Yan, Z. and Zong, L. (2020). Spatial prediction of housing prices in beijing using machine learning algorithms, *Proceedings of the 2020 4th high performance computing and cluster technologies conference & 2020 3rd international conference on big data and artificial intelligence*, pp. 64–71.

Zhou, H., Zhang, J., Zhou, Y., Guo, X. and Ma, Y. (2021). A feature selection algorithm of decision tree based on feature weight, *Expert Systems with Applications* **164**: 113842.