**House Price Prediction: Survey of Data Mining Methods and Regression Techniques**

***Radhika Reddy Chintareddy, Kusala Nagamani Reddy***

*The Graduate School, Clarkson University, Potsdam, NY, United States, Applied Data Science*

**ABSTRACT**

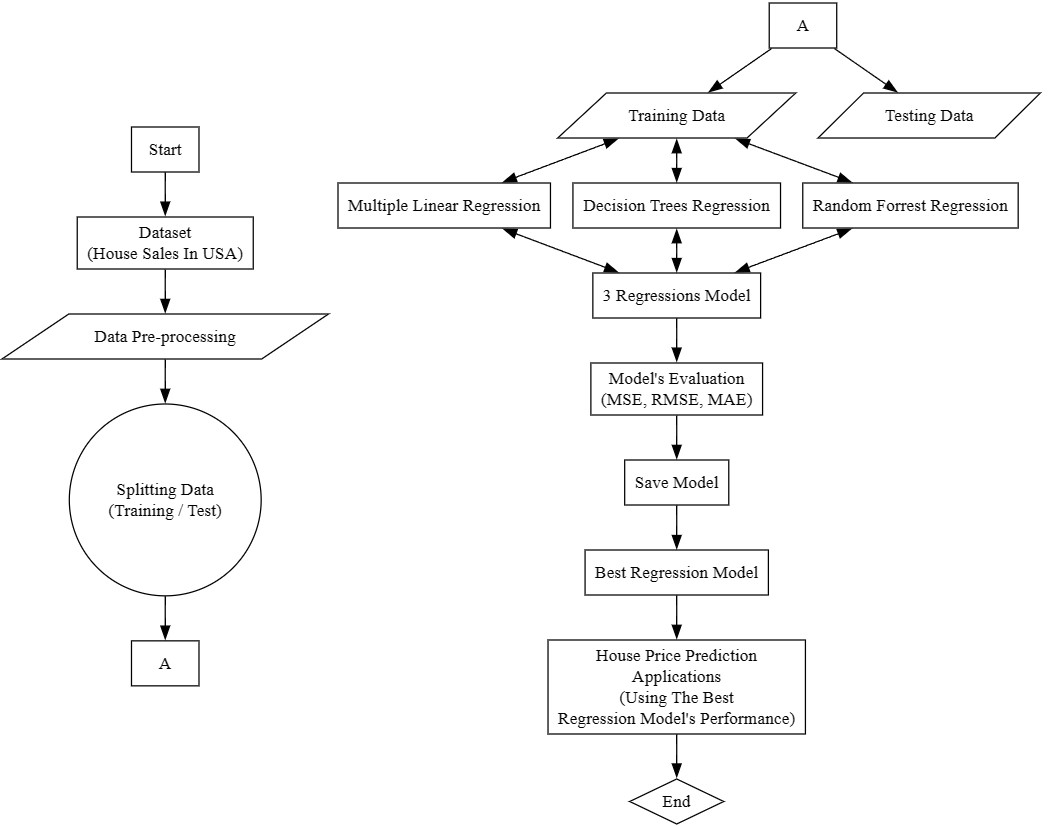
The Real Estate industry is dynamic in terms of the prices fluctuating regularly. As accurate house prices allow better-informing parties in the real estate market, improving housing policies and real estate appraisal, a comprehensive overview of house price prediction strategies is valuable for both research and society. Although conventional methods and traditional input data remain predominant, house price prediction research is slowly adopting more advanced techniques and innovative data sources. The project mainly focuses on predicting the real-valued prices for the places and the houses by applying the appropriate ML algorithms. Algorithms like Linear regression and sklearn are used to effectively increase the accuracy. During model structure, nearly all data similarities and cleaning, outlier removal and feature engineering, dimensionality reduction, gridsearchcv for hyperparameter tuning, k-fold cross-validation, etc. are covered. Model performance was measured using evaluation matrices such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and also Mean Absolute Error (MAE). Adjusted R-squared of 0.701 (70% of variations explained by our model). Our model identified a strong correlation between waterfront houses, Sight (if the house has already been viewed), and price. Also, there is a noticeable disparity among the zip codes. The median house price ranges from $235,000 in 98002 up to $1,260,000 in 98039. Overall, this model can aid housing developers and the common population alike.

**Keywords:** Linear regression model, Python, Machine Learning, House Price, Decision Tree.

1. **INTRODUCTION**

The need for houses is not only used as a place to live, they are also often used as long-term investment instruments by investors to get additional income from these types of investments, especially for property entrepreneurs who certainly produce investments that can be considered quite promising. House is one of the main human needs besides clothing and food. In the Hierarchy of Needs, the house is one part of the Basic Needs which can be categorized into 2 parts, namely Physiological Needs and Safety Needs. If the seller of the house makes a mistake, the result will be that the house will be less accepted on the market, this can make the house sellers have a smaller possibility or even lose the opportunity to get the maximum profit from the sale. To minimize errors in setting the selling price of the house, the seller must be careful in determining the price, that’s because the selling price of the house is mostly rising and rarely goes down either in the short or long term. The main drive behind the project is a prediction of real estate prices to build the best house price prediction systems using machine learning algorithms with maximum accuracy and without any loss. Many factors have to be taken into consideration for predicting house prices and try to predict efficient house pricing for customers concerning their budget as well as also according to their priorities. So, we are creating a housing cost prediction model. The price of a house depends on many factors like Area, location, population, size and number of bedrooms & bathrooms given, condition, quality, square footage, etc.

The proposed model aims to create an accurate result by taking into consideration all different factors. For House price prediction one can use various prediction models (Machine Learning Models) like MLR, Decision tree regression,), Logistic regression, artificial neural network, etc. House pricing model is beneficial for buyers, property investors, and house builders. This model will be informative and knowledgeable for the entities related to real estate and all the stakeholders to evaluate the current market trends and budget-friendly properties. The model initially concentrated on the analysis of the attributes that influence the prices of the houses based on Multiple Linear regression only but later we introduced Decision trees and random Forest regressors as well to Improve the metrics. The model building starts with the dataset from a reliable source that is simple to use. A dataset was chosen for our house price prediction, which contains 21613 records of data and 23 features for training our model. Various machine-learning procedures can be used to forecast future values (**Figure 1**). In any case, it is required a model that can forecast future property estimations with greater accuracy and less error.



**Figure 1 : Project workflow diagram.**

1. **METHODS**
   1. **Description of Data and Preprocessing**

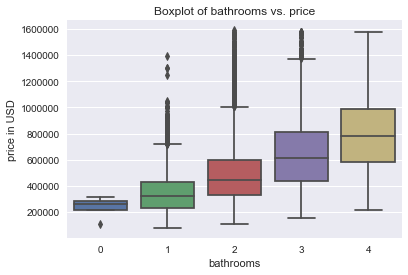
In the main flowchart (**Figure 1**), it can be seen that the model uses various regression techniques for predicting house prices. The house prices sales listing dataset in King County, USA will be used. House sales data in the dataset were collected and obtained from May 2014 to May 2015. The amount of data available is 21,613 data and has 23 features. The dataset is taken from the official Kaggle website. During Data Pre-processing, the dataset will first be analysed using graphical depictions and data checking using several data visualization techniques as well as Exploratory Data Analysis (EDA). Then the data will be checked for any duplicates, missing values, and outliers. Furthermore, based on the results of data visualization and analysis, the data that will be used for modeling needs to be cleaned first. The data that has gone through the pre-processing stages of the data will then be divided into 2 parts, where the first part of data will be used as training data, and the data in second part will be used as testing data.

* 1. **Data Visualization and Exploratory Data Analysis (EDA)**

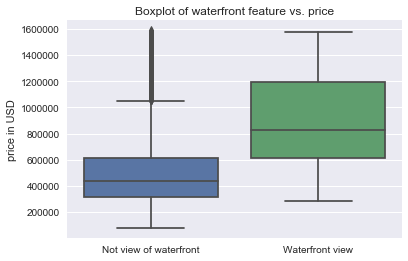
During the visualization and Exploratory data analysis (EDA) stages, features in the dataset except for ID, Date, Zipcode, Latitude, and Longitude, are analyzed for the distribution of the data to the target variable, which is the price of a house. The distribution analysis of the data was divided into 2 types, those are the analysis of continuous numerical features, and the second one is the analysis of discrete numerical features. Numerical continuous features, are: the size of living space, size of land area, flooring, and size of living space above ground level, Meanwhile, the features included in the Numerical Discrete Feature category are waterfront, view, condition, and grade.

The next step is to check whether the features in the dataset have normal data distribution or not. After checking, several variables/features have abnormal data distribution, this can be seen from the outliers. These features are price, size of living space, size of land space, size of living space above ground level, averages of living space sizes 15 closest neighbors, and averages of land space 15 closest neighbors. Checking the data distribution is also carried out from each of its features to the target variable, the price. From (**figure 3**) it can be observed that Whether the house is a waterfront house or not has a significant effect on the price of the house. Starting from continuous numerical features like the size of the living space and also the year the house was built have a big influence on the increase in the value of house prices. The age of the house was calculated from the year the house was built.

In features/variables that are numerically discrete, features that have a significant effect on increasing the selling price of a house are View, Waterfront, and Grade. The effect of those features on the increase in the selling price of a house shows an exponential increase in value. Other features such as the condition of the house also affect the increase in the selling price of the house, but the price increase tends to be quite stable with this feature, visualization using bar plots on discrete numerical features can be seen in (**Figure 2**).



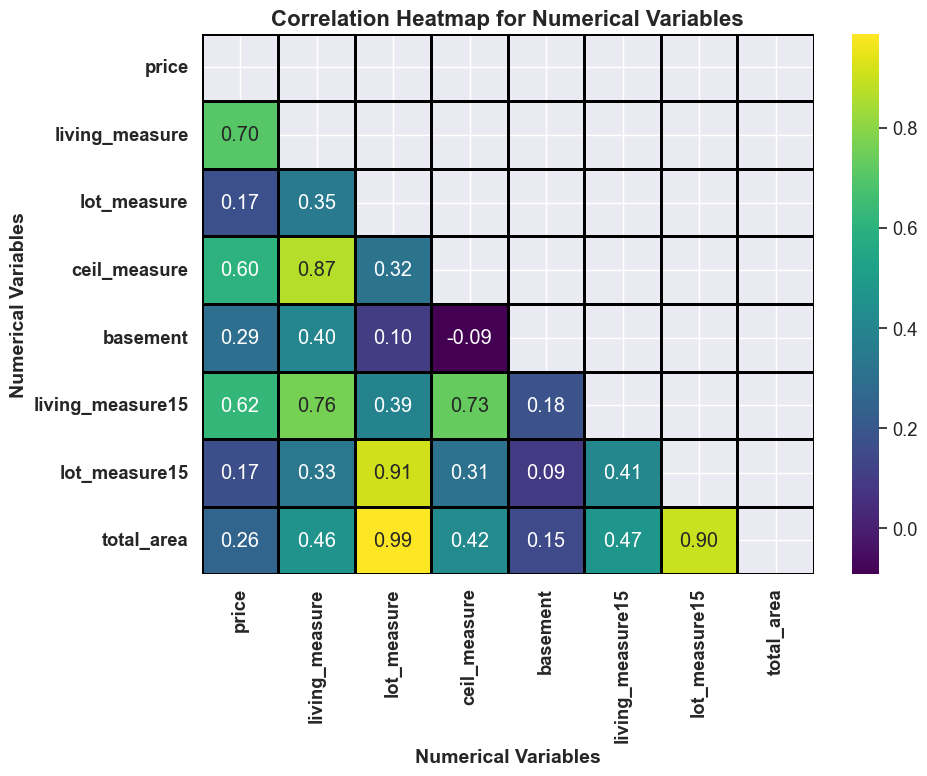
**Figure 2 : Price Vs Discrete variable**



**Figure 3 : Categorical feature plot**

**2.3 Data Cleansing**

Before diving into the analysis of the two types of features above, the dataset is first checked for the NULL values contained in the features, and cleaning the data is done by dropping several features that will not be used for the modeling process, these features are ID, Date, Zipcode, Latitude, and Longitude features. In this phase, cleaning and preparing the collected data for model training was done. Tasks such as handling missing values, removing outliers, and encoding categorical variables are performed as part of feature engineering. Feature selection techniques can be applied to identify the most relevant attributes for predicting house prices. Before building models, the data should be processed accordingly so that the models can learn the patterns more efficiently. Specifically, numerical values were standardized, while categorical values were one-hot-encoded. To check the multicollinearity Pearson correlation coefficient was used and the resulting heatmap is displayed. (**Figure 4**). It can be noticed that There is a good correlation between the Living measure feature and living measure 15 which says about neighbouring 15 properties and the same with lot measure prompting us to drop the living measure and lot measure. Also, after being processed, data splitting techniques such as Random sampling can be used to create training and testing datasets.

****

**Figure 4 : Pearson correlation heat map.**

**2.4 Clustering & Scaling the data**

Scaling is done on the data before clustering to ensure that each feature contributes equally to the clustering process, preventing biased results where certain features dominate due to their scale. Here two clusters are formed and analysis of the clusters reveals clear differences: Cluster 0 generally consists of higher-priced, larger, better-conditioned houses with superior views and furnishings, while Cluster 1 represents properties with lower values across these metrics. However, when regression analysis is applied to each cluster separately, it becomes apparent that the clusters only explain less than half of the overall variability in the dataset. Given below is (**Table 1**)Shows the results after performing linear regression on clusters separately. This suggests that the clusters might not fully capture the diversity present in the entire dataset, possibly due to the influence of categorical variables. Further investigation is needed to refine the clustering approach and improve its accuracy.

| **Cluster** | **Train RMSE** | **Test RMSE** | **Training Score** | **Test Score** |
| --- | --- | --- | --- | --- |
| High Cluster (Cluster 0) | 180325.283 | 179476.109 | 0.437610 | 0.432128 |
| Low Cluster (Cluster 1) | 128868.297 | 128001.968 | 0.421834 | 0.415300 |

**Table1 : Cluster regression results**

1. **RESULTS**

In the next step, the linear regression analysis conducted on the dataset yielded a model that explains approximately 65% of the variance in both the training and test data. While this demonstrates a moderate level of predictive ability, it suggests that the model may not fully capture the complexity of the data. Given is the regression equation to predict the price of the house based on the features. To further explore the importance of different features in predicting house prices, feature importance analysis was performed. The results of this analysis are presented in the accompanying feature importance plot, which highlights the relative significance of each feature in the regression model.

**\begin{equation}**

**\text{Price} = -14207.35 \times \text{room\\_bed} + 21372.84 \times \text{room\\_bath} + 25695.58 \times \text{ceil} + 140365.88 \times \text{coast} + 28660.38 \times \text{sight} + 20246.03 \times \text{condition} + 78870.96 \times \text{quality} + 84.86 \times \text{ceil\\_measure} + 114.64 \times \text{basement} + 58.63 \times \text{living\\_measure15} - 4.15 \times \text{lot\\_measure15} + 62563.28 \times \text{furnished} - 1.28 \times \text{total\\_area} + 2255.73 \times \text{house\\_age}**

**\end{equation}**

Moving forward, additional regression techniques will be explored to assess their efficacy in capturing the underlying patterns in the data. Specifically, decision trees, random forests, and artificial neural network (ANN) regressors will be employed. These algorithms offer the advantage of capturing non-linear relationships between features and target variables, which may better reflect the underlying structure of the dataset. By comparing the performance of these non-linear regression models with that of linear regression, we aim to identify the most suitable approach for predicting house prices in our dataset.

| **Regression Model** | **Train RMSE** | **Test RMSE** | **Training Score** | **Test Score** |
| --- | --- | --- | --- | --- |
| Linear Regression | 146602.06 | 146173.57 | 0.6592 | 0.6510 |
| Decision Tree Regressor | 8286.19 | 181869.34 | 0.9989 | 0.4598 |
| Random Forest Regressor | 49285.74 | 129987.63 | 0.9615 | 0.7240 |

**Table2 : Results before hyperparameter tuning**

| **Regression Model** | **Train RMSE** | **Test RMSE** | **Training Score** | **Test Score** |
| --- | --- | --- | --- | --- |
| Linear Regression | 146602.06 | 146173.57 | 0.6592 | 0.6510 |
| Decision Tree Regressor | 127342.60 | 142514.23 | 0.7429 | 0.6683 |
| Random Forest Regressor | 120945.44 | 133225.85 | 0.7680 | 0.7101 |

**Table2 : Results After hyperparameter tuning**

1. **DISCUSSION**

Hyperparameter tuning plays a critical role in optimizing the performance of regression models. For the Decision Tree Regressor, the best hyperparameters identified through grid search were a maximum depth of 20, minimum samples leaf of 30, and minimum samples split of 15. Similarly, for the Random Forest Regressor, optimal hyperparameters included a maximum depth of 10, maximum features of 6, minimum samples leaf of 3, and minimum samples split of 30, with 101 estimators.

This tuning process aims to find the best combination of hyperparameters that minimizes the model's error while avoiding overfitting. The results demonstrate that after hyperparameter tuning, both the Decision Tree and Random Forest regressors outperform the initial Linear Regression model in terms of both training and test scores. Specifically, the Random Forest Regressor achieved the lowest test RMSE and highest test score, indicating improved generalization performance compared to the other models. These findings underscore the importance of hyperparameter tuning in optimizing regression models for predictive accuracy and generalizability in real-world applications.

1. **CONCLUSIONS**

In conclusion, our comprehensive analysis of housing data within King County has yielded valuable insights for real estate market strategies. We observed that waterfront properties and upscale neighborhoods like Medina, Clyde Hill, and Mercer Island command significantly higher prices. Additionally, attributes such as house grade and bedroom count emerged as strong predictors of price. Timing-wise, launching campaigns in March/April aligns with peak sales periods in Q2.

Our modeling efforts, particularly the hyperparameter tuning of regression algorithms, significantly enhanced predictive performance. Notably, the Random Forest Regressor exhibited superior performance without signs of underfitting, underscoring its effectiveness in capturing complex relationships within the data. Furthermore, future enhancements could explore alternative regression algorithms like Ridge, Lasso, Bagging, and Boosting, potentially yielding further improvements in model accuracy.

In future endeavors, incorporating additional datasets on commuting times, income distributions, longer-term trends, and school rankings could provide richer insights and refine our predictive models. Ultimately, our findings lay the groundwork for developing more tailored and effective real estate market strategies, empowering stakeholders to make informed decisions and capitalize on emerging opportunities in the housing market.

1. **REFERENCES**

Frontiers | Differential Impact of COVID-19 Risk Factors on Ethnicities in the United States (frontiersin.org)

A Comprehensive Guide on Hyperparameter Tuning and its Techniques (analyticsvidhya.com)

Regression Analysis: Step by Step Articles, Videos, Simple Definitions (statisticshowto.com)

Regression analysis—ArcGIS Insights | Documentation

Subhani Shaik, Uppu Ravibabu. Classification of EMG Signal Analysis based on Curvelet Transform and Random Forest tree Method. Paper selected for Journal of Theoretical and Applied Information Technology (JATIT). 95. 15.

Shiva Keertan J, Subhani Shaik. Machine Learning Algorithms Prediction, for International Oil Journal Price of Innovative Technology and Exploring Engineering. 8(8). 16.

KP Surya Teja, Vigneswar Reddy and Subhani Shaik, Flight Delay Prediction Using Machine Learning Algorithm XGBoost, Jour of Adv Research in Dynamical & Control Systems. 11(5).